Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in

Scienze Statistiche

Analisi statistica delle reti: un'applicazione alla mobilità dei docenti italiani fra diversi Atenei

Relatore Prof. Adriano Paggiaro
Dipartimento di Scienze Statistiche

Correlatore Dott.ssa Caterina De Bacco
Santa Fe Institute, New Mexico

Laureanda Sara Ferrari
Matricola N 1100334

Anno Accademico 2016/2017

Sometimes you gotta get through your fear
to see the beauty on the other side.

Indice

Elenco delle figure

Elenco delle tabelle

Capitolo 1

Introduzione

L'obiettivo dell'analisi statistica delle reti è la modellazione di complesse strutture di dati, nei quali l'interesse si sposta dallo studio dell'unità statistica allo studio delle connessioni tra diverse unità. L'aspetto rilevante sono quindi le relazioni, che possono essere di vicinanza, interazione, scambio etc. Questa materia sta trovando negli ultimi anni una sempre più vasta applicabilità in contesti quali tecnologia, informazione, biologia e sociologia.

Un'interessante applicazione nell'ambito delle scienze sociali riguarda la mobilità dei lavoratori. Negli ultimi anni, il concetto di mobilità è spesso al centro del dibattito pubblico e politico italiano, per i cambiamenti che stanno avvenendo nel mercato del lavoro, ed è perciò un tema molto attuale. Anche in ambito accademico l'essere "mobili", per i laureati che desiderano intraprendere una carriera in tale ambito e per i docenti, ha assunto negli ultimi tempi una connotazione diversa, generata dal fatto che in tutti i campi lavorativi c'è la tendenza a contratti più flessibili e in generale a tempo determinato. Ciò sta portando a una modifica della visione classica in questo campo, per la quale il professore universitario è considerato in Italia un "posto fisso". Proprio in merito a questa classica concezione, esiste una sostanziale differenza tra le culture del lavoro in ambito accademico nelle diverse parti del mondo.

L'articolo da cui prende spunto questo lavoro è scritto da Clauset, Arbesman e Larremore (2015), i quali studiano la mobilità all'interno di diversi settori accademici negli Stati Uniti attraverso una modellazione di reti. Negli Stati Uniti, la situazione generale è di una spiccata volontà di muoversi degli individui dall'ateneo

dove si laureano o dove conseguono il Ph.D. ai successivi posti di lavoro e si può pensare che questo derivi da una particolarità culturale.

Lo studio portato avanti in questa tesi nasce dall'idea che ci sia in questo ambito una diversità tra la cultura americana e quella italiana. L'obiettivo è quello di valutare diverse caratteristiche della rete degli atenei italiani per quanto riguarda due ambiti quali la Statistica e la Fisica sperimentale che, seppur vicini da un punto di vista scientifico, presentano alcune differenze rilevanti.

Questo lavoro si propone due obiettivi distinti: la *community detection* degli atenei italiani e la creazione di un *ranking* tra gli stessi. Il primo aspetto è di tipo descrittivo e interpretativo: si vogliono identificare dei gruppi di atenei simili dal punto di vista di formazione e assunzione dei docenti, valutando in merito a questo se geograficamente si può dire che ci sia mobilità tra i docenti universitari nello spostamento dall'ateneo dove si studia a quello dove si lavora. Il secondo deriva dallo studio dei pattern di spostamento dei docenti tra gli atenei; è infatti possibile creare un ordinamento degli atenei generato dal prestigio e dallo status sociale degli stessi. In merito a questo, può esserci un ulteriore obiettivo esterno per quanto riguarda da un lato l'assegnazione dei fondi da parte dello Stato, che sono erogati in modo proporzionale alla qualità della ricerca nelle università, dall'altro la scelta dell'ateneo da potenziali fruitori dei servizi dello stesso.

La trattazione si sviluppa come segue.

Nel secondo capitolo viene introdotto il contesto teorico e modellistico alla base dell'analisi statistica delle reti, a partire dalla struttura matematica necessaria a spiegare questa tipologia di dati, passando per diverse specificazioni di una delle categorie di modelli usati in questo ambito, gli *Stochastic Blockmodels*, fino ad arrivare all'introduzione del modello utilizzato per l'analisi dei dati. Viene inoltre presentato un algoritmo ideato per estrarre un ranking dalla rete creata.

Nel terzo capitolo è descritto il modo in cui sono stati raccolti i dati, per metà provenienti da un precedente lavoro e per metà raccolti *ex novo* per questa tesi. Vengono inoltre presentate alcune analisi di tipo esplorativo che sono fondamentali per una prima lettura delle informazioni contenute nei dati. Questa tipologia di analisi è stata portata avanti parallelamente da un punto di vista statico e dinamico: inizialmente si guarda una fotografia dei dati divisi nelle due unità fondamentali di analisi, gli atenei e i docenti; nella seconda parte, invece, il problema viene affrontato in una prospettiva dinamica e viene descritta la rete costruita dall'insieme

delle due unità di analisi.

Nel quarto capitolo si applicano il modello e l'algoritmo spiegati nella seconda sezione, con l'obiettivo da un lato di descrivere i dati tramite comunità di atenei dal comportamento simile nella formazione e nell'assunzione di docenti nello specifico settore disciplinare; dall'altro, si cerca di stilare una classifica di atenei che esprima un ordinamento sulla base del prestigio degli stessi, grazie all'informazione racchiusa nella rete.

Si conclude con il quinto capitolo, dove si sono riassunte considerazioni di ordine generale volte a fornire un'analisi critica dei risultati, cercando di evidenziare pregi e difetti degli stessi.

Capitolo 2

Modelli per l'analisi delle reti

2.1 Dai grafi ai modelli per reti

La struttura matematica adatta a spiegare la particolare tipologia di dati, dove l'interesse è sì nelle unità ma anche e soprattutto nelle connessioni o relazioni tra di esse, è il grafo. Formalmente, un grafo $G = (V, E)$ è la struttura formata da un insieme V di *nodi* o *vertici*, le unità, e un insieme E di *archi*, le relazioni che intercorrono tra i nodi. Nella sua forma base è chiamato *grafo semplice* e non contiene archi le quali estremità sono connesse allo stesso nodo (chiamati *auto-archi* o *loop*) e coppie di vertici con archi multipli tra di essi. Un grafo che presenta entrambe queste caratteristiche è chiamato *multigrafo*. Inoltre, un grafo G in cui gli archi hanno una direzionalità, dove cioè un'eventuale relazione tra i nodi $\{u, v\}$ è quantitativamente diversa dalla relazione $\{v, u\}$ con $u, v \in V$, è chiamato *grafo direzionato* od *orientato*.

La struttura di interesse per i dati che si utilizzano in questa tesi presenta tutte le caratteristiche appena menzionate. Si possono quindi definire alcuni concetti specifici in questo ambito che saranno poi utili per la descrizione dei dati e della rete che li descrive.

Il numero di nodi $N = |V|$ e il numero di archi $N_e = |E|$ sono chiamati rispettivamente *ordine* e *grandezza* del grafo G.

Un grafo $H = (V_H, E_H)$ è un *sottografo* di un altro grafo $G = (V_G, E_G)$ se $V_H \subseteq V_G$ e $E_H \subseteq E_G$; questo concetto sarà utile in relazione all'obiettivo di

suddividere il grafo completo in gruppi, che saranno quindi dei sottografi.

La connettività di un grafo è facilmente descrivibile tramite la *matrice di adiacenza* dello stesso e tramite alcuni concetti a essa correlati. La matrice di adiacenza A di un grafo semplice è una matrice quadrata di dimensione $N \times N$ i cui elementi:

$$A_{ij} = \begin{cases} 1, \text{ se } \{i,j\} \in E \\ 0, \text{ altrimenti} \end{cases} .$$

Nel caso dei multigrafi la matrice è facilmente generalizzabile, associando un valore naturale pari al numero di connessioni tra due nodi al posto del semplice valore unitario e permettendo la presenza di valori diversi da zero sulla diagonale. Il nome *adiacenza* deriva dalle definizioni di archi e nodi adiacenti: in particolare, due nodi sono adiacenti se sono connessi da un arco mentre due archi sono adiacenti se hanno un estremo in comune nello stesso nodo. Inoltre, se un nodo v è un estremo di un arco e si dice che e è incidente a v e si può costruire la *matrice di incidenza* seguendo questa definizione.

Da quanto detto discende anche il concetto di *grado* di un vertice v, d_v, definito come il numero di archi incidenti a v. Si noti che, nel caso di interesse dei grafi direzionati, il grado del vertice si può ripartire in *grado in uscita* (d_v^{out}) e *grado in entrata* (d_v^{in}), i quali contano rispettivamente il numero di archi che partono da o che puntano verso un vertice.

É utile, per concludere, discutere il concetto di movimento all'interno del grafo, che viene espresso tramite la definizione di *percorso* o *path*, il quale è una sequenza alternata di vertici e archi $\{v_0, e_1, v_1, e_2, ..., v_{l-1}, e_l, v_l\}$ che definisce il percorso di lunghezza l dal vertice v_0 al vertice v_l. Per conoscere la distanza tra due generici vertici, è comune utilizzare il concetto di *percorso più corto* che è possibile trovare nel grafo per connettere i due vertici. La distanza così definita viene spesso chiamata *distanza geodetica*. Per una visione generale e più approfondita riguardo la teoria dei grafi come base per lo studio statistico delle reti si rimanda a Kolaczyk (2009).

Dopo la definizione di questi aspetti applicabili a un singolo grafo risulta intuitiva l'estensione al concetto di *grafo multilivello*. Un *grafo multilivello* (che può essere diretto o indiretto, semplice o multigrafo, a seconda dei grafi che lo compongono) è la collezione di L grafi

$$G^{(\alpha)} = \{V, E^{(\alpha)}\}, \ \alpha = 1, ..., L$$

dove l'insieme di vertici è comune a tutti gli L livelli (o strati), ma l'insieme di archi varia. In questo modo si può descrivere una situazione in cui tra gli stessi nodi sono presenti diverse tipologie di connessioni, rappresentate da archi in diversi livelli. Ogni livello avrà la propria matrice di adiacenza $A^{(\alpha)}$ di dimensione $N \times N$, dove l'elemento $A^{(\alpha)}_{i,j}$ in un grafo semplice è pari a 1 se esiste un arco tra i e j nel livello α, o analogamente un arco di tipo α. In alternativa, si può pensare all'insieme delle matrici $A^{(\alpha)}$ come a un tensore A di dimensione $N \times N \times L$, come si vedrà nella sezione 2.4.

Una volta compresa la struttura matematica delle reti, l'interesse si sposta nel trovare dei modelli statistici adatti a rappresentarla e descriverla.

Con modello per rete viene indicata una collezione

$$\{\mathbb{P}_\theta(G),\ G \in \mathcal{G} : \theta \in \Theta\},$$

dove $\mathcal{G}$ è un insieme di possibili grafi e $\mathbb{P}_\theta$ è la distribuzione di probabilità su $\mathcal{G}$, con θ vettore dei parametri contenuti nello spazio parametrico Θ. Da questa semplice formulazione si può pensare a un numero molto ampio di modelli possibili, sulla base di come viene scelta la distribuzione di probabilità.

Associare un modello probabilistico a una struttura matematica come il grafo trova la sua motivazione nell'esigenza di descrivere dei dati dai quali si vuole estrarre tutta l'informazione contenuta fornendo allo stesso tempo un background teorico-probabilistico. Uno degli obiettivi principali che gli studiosi della materia cercano di perseguire è la ricerca di gruppi nei dati, chiamata comunemente *community detection*, spiegata dalla tendenza di molte reti reali di dividersi naturalmente in gruppi. Come brevemente citato in precedenza, e come si vedrà nello specifico nell'applicazione che segue, nel caso in esame si vogliono trovare gruppi di atenei universitari che si comportano in modo simile rispetto al pattern di formazione e reclutamento dei docenti.

Nei seguenti paragrafi si cerca di contestualizzare il modello utilizzato nell'analisi dei dati, partendo dalla classe ampia di modelli chiamati *Stochastic Blockmodels* e delineandone caratteristiche di base e successive estensioni e modifiche.

2.2 Gli Stochastic Blockmodels

Gli articoli pionieri della teoria che traduce i grafi da struttura matematica a oggetto probabilistico sono quelli di Erdös e Rényi (1959, 1960), i quali hanno studiato le proprietà dei grafi *bernoulliani* e i modelli a essi collegati. Data la difficile applicazione di questi modelli teorici ai dati reali, si è successivamente iniziato ad ampliare la scelta delle possibili modellazioni e in particolare in letteratura sono stati proposti i *Blockmodels*, generalizzati successivamente in un contesto probabilistico in *Stochastic Blockmodels* (si veda ad esempio Fienberg e Wasserman (1981) e Holland, Laskey e Leinhardt (1983)).

L'obiettivo dei modelli a blocchi è quello di partizionare l'insieme dei vertici in gruppi (o appunto *blocchi*) in modo che questa divisione colga nel modo migliore la struttura di connessioni che esiste nella rete. Si parla di *equivalenza strutturale* riferendosi alla somiglianza tra due nodi che appartengono allo stesso gruppo nel modo in cui si relazionano con altri nodi e in generale con la rete.

Inizialmente, questa tipologia di modelli adottava un approccio di tipo deterministico permettendo di trovare la permutazione dei nodi che più approssimava la struttura a blocchi. Naturalmente, questi procedimenti mancavano di una base statistica ed erano di natura per lo più esplorativa.

Per questo motivo Fienberg e Wasserman (1981) e altri autori hanno esteso questi concetti inscrivendoli in un contesto probabilistico. Uno *Stochastic blockmodel* si può definire come una famiglia di distribuzioni di probabilità per grafi, i quali vertici sono suddivisi in gruppi chiamati blocchi, che hanno la proprietà di contenere nodi *stocasticamente equivalenti* tra loro, nel senso che per ogni nodo all'interno di un gruppo, condizionatamente all'appartenenza allo stesso, la probabilità di avere relazioni con altri è equivalente. La distribuzione di probabilità è poi invariante a permutazioni di nodi all'interno dei gruppi (Snijders e Nowicki 1997).

Si può fare una suddivisione di questi ultimi modelli in base alla conoscenza o meno dei blocchi a priori. Il contesto dove si colloca questo lavoro è quello in cui non si conosce a priori la natura e la numerosità dei gruppi di nodi e quindi si parla di *a posteriori blockmodeling*, procedura che si può inserire quindi in ambito non supervisionato.

Un lavoro molto importante è quello di Snijders e Nowicki (1997), i quali,

a partire dalla struttura illustrata nel paragrafo 2.1 definiscono un blockmodel casuale, specificando i seguenti aspetti:

- Ogni elemento della matrice di adiacenza A, A_{ij}, è una variabile casuale tale che per $i < j$ le variabili sono statisticamente indipendenti; inoltre $A_{ij} \equiv A_{ji}$ e $A_{ii} \equiv 0$ (si noti che queste caratteristiche definiscono un grafo semplice).

- Si assume che esista una partizione dell'insieme di nodi in blocchi tale che per ogni vertice i, j, h, se $i \neq j \neq h$ e se i e h appartengono allo stesso gruppo, allora A_{ij} e A_{hj} sono identicamente distribuite.

Definizione. *Un blockmodel casuale è una famiglia di distribuzioni di probabilità per un grafo semplice suddiviso in blocchi (chiamato* grafo colorato*) G con nodi $\{1, ..., N\}$ e gruppi o colori $\mathcal{C} = \{1, ..., K\}$, definito come segue.*

1. *L'insieme dei parametri è costituito dal vettore $\Theta = (\theta_1, ..., \theta_K)$ delle probabilità relative ai gruppi e dalla matrice $\eta = (\eta_{kl})_{1 \leq k \leq l \leq K}$ di probabilità relative agli archi tra coppie di nodi, che dipendono solo dal gruppo a cui appartengono i nodi.*

2. *Il vettore dei colori dei vertici consiste nelle variabili casuali i.i.d. $(X_i)_{i=1}^N$, dove $\mathbb{P}(X_i = k) = \theta_k$ per $k = 1, ..., K$.*

3. *Condizionatamente ai colori dei vertici X_i, gli archi A_{ij} sono indipendenti con $A_{ij} \sim Bernoulli(\eta_{X_i, X_j})$. Detto in altre parole, la distribuzione della struttura relazionale può essere modellata condizionatamente al vettore dei colori e in particolare ogni A_{ij} tramite i suoi attributi X_i e X_j.*

Se (A, X) è quindi ciò che rappresenta un grafo colorato G, la funzione di probabilità è data da

$$\mathbb{P}(G; \Theta, \eta) = \theta_1^{n_1} \cdots \theta_K^{n_K} \prod_{1 \leq k \leq l \leq K} \eta_{kl}^{e_{kl}} (1 - \eta_{kl})^{n_{kl} - e_{kl}} \tag{2.1}$$

dove $n_k = \sum_{i=1}^N I(x_i = k)$ conta il numero di vertici della rete di colore k,

$$e_{kl} = \frac{1}{1 + \delta_{kl}} \sum_{1 \leq i \neq j \leq N} A_{ij} I(x_i = k) I(x_j = l)$$

denota il numero di archi che hanno un vertice di colore k e l'altro di colore l, e

$$n_{kl} = \begin{cases} n_k n_l & se\ k \neq l \\ \binom{n_k}{2} & se\ k = l \end{cases} \qquad \delta_{kl} = \begin{cases} 0 & se\ k \neq l \\ 1 & se\ k = l \end{cases}.$$

Si potrebbe intuitivamente pensare, come poi si riscontra in molte applicazioni sui dati reali, che le relazioni più frequenti in una rete siano all'interno dei blocchi piuttosto che tra blocchi diversi (per esempio le relazioni di amicizia tra gruppi di individui con un simile stile di vita). In questo caso le probabilità η_{kl} saranno generalmente più piccole rispetto a η_{kk}. É comunque possibile avere degli esempi di reti in cui i valori diagonali η_{kk} sono più piccoli di quelli fuori della diagonale, come per esempio se si parla di attrazione sessuale e i blocchi sono definiti dal genere in una popolazione per la maggioranza eterosessuale. Si chiamerà questo aspetto della rete *struttura* ed essa sarà caratterizzata dalla presenza o assenza di una particolare proprietà delle reti chiamata *omofilia*, o *assortatività*, che è appunto la tendenza a legarsi tra soggetti simili per un qualche aspetto.

Come detto, l'approccio di Snijders e Nowicki (1997) prevede solamente la modellazione di grafi semplici in cui, quindi, non sono presenti auto-archi o archi multipli e in particolare nell'articolo viene esplicitato il caso in cui i gruppi sono due. Gli stessi autori hanno poi generalizzato questo approccio permettendo la presenza di archi direzionati e pesati (possono assumere diversi valori) e di un numero arbitrario di classi (Nowicki e Snijders 2001). Ci si avvicina quindi alla situazione di interesse ma con ancora degli aspetti critici, come per esempio la non considerazione degli auto-archi. Il vettore degli attributi X che specifica la struttura a classi è considerato *latente* e in generale la struttura è chiamata *struttura relazionale colorata*.

Bisogna per esempio generalizzare la matrice η delle probabilità degli archi in quanto in questo nuovo modello non si è più in un caso dove l'arco può essere semplicemente presente o assente, ma può assumere un valore naturale positivo che specifica quanti archi sono presenti tra due nodi.

Si possono derivare a questo punto la distribuzione condizionata delle relazioni A dato il vettore di colori X e successivamente la distribuzione congiunta (A, X) che definisce lo *stochastic blockmodel*. Per evitare un abuso di notazione, per la definizione completa del modello si rimanda al paragrafo 2.3, dove, seguendo l'approccio di Ball, Karrer e Newman (2011), si fa un ulteriore passo avanti nella

modellazione e si mettono le basi per la specificazione finale del modello usato in questa tesi.

Per quanto riguarda la stima di questi due modelli di base si fa notare come, per il primo modello presentato, siano state proposte più di una modalità di stima. In particolare, tramite massima verosimiglianza e l'algoritmo EM (Dempster, Laird e Rubin 1977) oppure tramite stima bayesiana con il Gibbs sampling (Geman e Geman 1984), il quale si è dimostrato essere più flessibile. Per questo motivo, nel modello generalizzato viene utilizzata solo la procedura tramite Gibbs sampling. Maggiori dettagli per quanto riguarda il processo di stima saranno forniti nei paragrafi successivi.

2.3 Un modello per comunità miste di arco

Un aspetto caratterizzante dei precedenti modelli è il fatto di considerare ogni nodo appartenente a uno e uno solo gruppo e quindi che i blocchi siano disgiunti tra loro. Questo potrebbe essere limitante in situazioni reali dove un'unità potrebbe appartenere a gruppi diversi, come per esempio nelle reti sociali, dove una persona può appartenere a diversi gruppi di conoscenze (amici, parenti etc.). In questo caso si parla di gruppi sovrapposti in quanto hanno almeno un'unità nella loro intersezione. Questo sembra sensato nella situazione in esame perché si cerca di inferire una struttura di gruppo di cui non si hanno informazioni a priori, che quindi potrebbe essere formata da gruppi tra loro sovrapposti. Si ha quindi più flessibilità, imponendo meno restrizioni.

Un modello che tenga conto di questa possibilità è stato formalizzato da Airoldi et al. (2008), dove a ogni nodo viene associato un vettore di probabilità di appartenenza ai vari gruppi mantenendone la natura latente. In questo modello è più probabile che due vertici siano connessi se hanno più di un gruppo in comune e questo implica che l'area sovrapposta tra due comunità abbia una densità media maggiore di archi rispetto all'area di una singola comunità (Ball, Karrer e Newman 2011).

Per avere meno restrizioni riguardo la struttura di sovrapposizione tra i gruppi, Ball, Karrer e Newman (2011) hanno sviluppato un procedimento che si basa sul concetto di *comunità di arco*. L'idea (sviluppata anche da altri autori) molto semplice alla base è quella per cui si considera la presenza di diversi tipi di archi

in una rete. Se si raggruppano gli archi sulla base del loro comportamento nel collegare i nodi, innanzitutto si possono dedurre le comunità dei nodi, semplicemente vedendo a quali archi sono connessi, e in secondo luogo è ovvia la possibilità di appartenenza di un nodo a più gruppi che è data dal fatto che un nodo può essere connesso a diversi tipi di archi.

Un altro aspetto interessante e particolare del lavoro di questi autori è l'utilizzo di un modello generativo invece di un approccio basato su una funzione obiettivo. Essendo inoltre un modello generativo per comunità di arco, a differenza di uno generativo per comunità di nodo dove prima vengono assegnati i nodi e successivamente gli archi, l'assegnazione di archi e nodi viene fatta simultaneamente.

Si noti che nemmeno questo modello è utilizzabile per i dati in esame perché il grafo non è diretto né pesato, anche se generalizzato al caso dei multigrafi. Si può affermare, però, che si tocca un altro aspetto fondamentale necessario per il problema di questa tesi e cioè le comunità non disgiunte. Si vedrà poi che i modelli spiegati fino ad ora sono la genesi del modello che effettivamente viene utilizzato e che presenta tutti gli aspetti necessari per descrivere i dati a disposizione.

Riprendendo la notazione usata fino a questo punto, il modello genera una rete con un numero N di vertici e archi non direzionati divisi in K comunità. A questo proposito, conviene modificare la formulazione dei parametri θ in quanto diventano dei vettori di lunghezza K con entrata θ_{iz} che rappresenta la propensione del vertice i di avere archi a esso collegati di colore, o appartenenti al gruppo, z.

In particolare $\theta_{iz}\theta_{jz}$ è il numero atteso di archi di colore z tra i vertici i e j. Si assume che la presenza degli archi sia distribuita come una Poisson di media questo valore, chiarendo quindi come viene modellata la presenza di più archi tra una coppia di nodi. Inoltre, vengono permessi anche i *loops* ed essi hanno valore atteso pari a $\theta_{iz}\theta_{iz}$. Il numero atteso totale di archi di qualsiasi colore tra due nodi si ottiene tramite la somma:

$$\sum_{z=1}^{K} \theta_{iz}\theta_{jz} \tag{2.2}$$

e la rispettiva per gli auto-archi.

In questo modello è naturale pensare alla creazione delle comunità di arco perché se due vertici i e j hanno valori alti di un certo θ_{iz} e il corrispondente θ_{jz},

significa che hanno un'alta probabilità di essere connessi a un arco di colore, o tipo, z e quindi appartenere a quel gruppo.

La probabilità di generare un grafo G con matrice di adiacenza A è pari al prodotto della probabilità relativa agli archi fuori dalla diagonale e la probabilità relativa agli auto-archi:

$$\mathbb{P}(G|\Theta) = \prod_{i<j} \frac{(\sum_z \theta_{iz}\theta_{jz})^{A_{ij}}}{A_{ij}!} exp\big(-\sum_z \theta_{iz}\theta_{jz}\big) \times \prod_i \frac{(\sum_z \theta_{iz}\theta_{iz})^{A_{ii}}}{A_{ii}!} exp\big(-\sum_z \theta_{iz}\theta_{iz}\big). \quad (2.3)$$

Si nota, a differenza della formulazione 2.1 vista in precedenza, che il parametro in questo caso sono solo i vettori θ.

Prendendo la log-verosimiglianza e usando la disuguaglianza di Jensen, si arriva naturalmente all'uso dell'algoritmo EM (Dempster, Laird e Rubin 1977) per la stima del modello, in quanto la massimizzazione della log-verosimiglianza può essere fatta ottimizzando i parametri Θ e una loro trasformazione, iterativamente.

2.4 Un'estensione per grafi multilivello con comunità miste

Riassumendo quanto detto fin'ora, il filone di ricerca su cui ci si è posti è quello dei *Stochastic blockmodel* (Snijders e Nowicki 1997) che hanno però degli evidenti limiti per quanto è di interesse in questo lavoro: nelle forme base non permettono la modellazione di archi multipli e auto-archi. Una generalizzazione di questi modelli che permette queste forme particolari di collegamenti, è invece carente dal punto di vista della direzionalità degli archi e quindi permette solo un collegamento bidirezionale con matrice di adiacenza simmetrica, considerando però la possibilità di appartenenza a più comunità per ogni nodo (Ball, Karrer e Newman 2011).

Un modello costruito appositamente per tenere conto di tutti questi aspetti critici è quello di De Bacco et al. (2017), che si posiziona ancora in questo ambito teorico ma da un lato sviluppa gli aspetti carenti visti in precedenza e dall'altro fornisce la possibilità di stimare una rete con più livelli.

Come sarà chiarito nel prossimo capitolo, i dati a disposizione hanno due possibili modi di considerare l'istruzione degli individui: a livello di laurea o Ph.D. Questo modello permette di considerare sia un caso di un unico livello, sia il caso più generale multi-livello in cui si usa l'informazione su entrambi i titoli di studio.

L'assunzione alla base di questa tipologia di modelli è che la rete abbia una struttura di gruppo implicita che si manifesta attraverso l'insieme delle connessioni tra i nodi nei diversi strati, i quali hanno ognuno una diversa percentuale di informazione e di errore (Paul e Chen 2015). Si può estrarre informazione da una struttura di questo tipo sia aggregando i livelli dopo aver stimato la struttura di gruppo in ognuno, sia stimando la struttura nel suo complesso cercando di sfruttare tutta l'informazione che proviene da ogni livello separatamente.

Gli obiettivi di un modello così complesso sono molteplici: innanzitutto, inserendosi in un contesto di *community detection*, si cerca di descrivere la rete tramite la sua struttura in gruppi latente. Secondo, vista la presenza di queste diverse tipologie di connessioni tra i nodi, si cerca di capire la struttura di dipendenza tra gli strati e in relazione a questo se ci sono livelli non informativi e quindi se solo alcuni sono necessari per spiegare le comunità. Infine, si vuole un modello di tipo generativo in modo da poterlo usare per fare previsioni basate sulla densità anche sui dati mancanti, per avere quindi una stima anche di archi che nei dati raccolti non esistono (De Bacco et al. 2017).

La rete è formata da N nodi e ha L livelli. Come detto nel paragrafo 2.1, la rete avrà più di una matrice di adiacenza e in particolare si può pensare ad A come a un tensore $N \times N \times L$. Un elemento di una singola matrice sarà denotato con $A_{ij}^{(\alpha)}$ e conterrà il numero di archi tra i e j di tipo α.

Ogni nodo può appartenere a più di una comunità e la sua struttura di appartenenza è descritta da un vettore K-dimensionale, assumendo la presenza implicita di K gruppi sovrapposti nella rete. Data la direzionalità della rete, è interessante valutare due tipologie di appartenenza ai gruppi: un nodo appartiene a una certa *incoming community* sulla base degli archi puntati verso di esso, e appartiene a una *outgoing community* sulla base di quelli che da esso partono. Si denotano questi due vettori relativi al nodo i, che nel caso non direzionato sono uguali, con u_i e v_i.

Un aspetto molto interessante di questo modello è la possibilità per ogni strato di avere una sua specifica struttura. Come già menzionato in precedenza, spesso la struttura delle relazioni in una rete è di tipo *assortativo* ma può valere anche l'op-

posto; mentre molti modelli impongono una determinata struttura a tutti i livelli della rete, in questo caso alla struttura viene permesso di variare, introducendo quindi una matrice strutturale $w^{(\alpha)}$ che è diversa da livello a livello, di dimensione ogni volta $K \times K$.

Il numero atteso di archi nel livello α da i a j è quindi

$$M_{ij}^{(\alpha)} = \sum_{k,l=1}^{K} u_{ik} v_{jl} w_{kl}^{(\alpha)}, \tag{2.4}$$

mostrando una chiara analogia con la (2.2) ma allo stesso tempo distanziandosene nettamente per l'uso di due vettori diversi di appartenenza che sono dovuti dal fatto che dal nodo i parte l'arco, mentre il nodo j lo riceve. Altro aspetto differente è l'aggiunta della matrice strutturale, che nei modelli precedenti non veniva inserita, considerandola semplicemente assortativa. Si noti, però, che l'appartenenza ai gruppi viene condivisa da tutti i livelli e quindi anche se l'informazione è portata da ogni livello in modo autonomo, un nodo ha solo due vettori di appartenenza.

Per ogni i, j e α la realizzazione di cui disponiamo $A_{ij}^{(\alpha)}$ deriva da una Poisson con media $M_{ij}^{(\alpha)}$ (analogamente al modello 2.3 nel caso di un livello):

$$\mathbb{P}(G|\Theta) = \prod_{i,j=1}^{N} \prod_{\alpha=1}^{L} \frac{e^{-M_{ij}^{(\alpha)}} (M_{ij}^{(\alpha)})^{A_{ij}^{(\alpha)}}}{A_{ij}^{(\alpha)}!} \tag{2.5}$$

dove si usa Θ per scrivere in modo compatto tutti i parametri u_{ik} , v_{jl} e $w_{kl}^{(\alpha)}$, che sono in totale $2NK + K^2L$.

2.4.1 Stima tramite l'algoritmo EM

Data una rete osservata, l'obiettivo del modello proposto da De Bacco et al. (2017) è quello di stimare contemporaneamente le appartenenze dei nodi e la struttura dei livelli della rete.

Ponendo una distribuzione a-priori nei parametri Θ si potrebbe porre il modello in un contesto bayesiano e fare inferenza sulla a-posteriori; non facendolo, si può lavorare sulla verosimiglianza (2.5) o in modo analogo sulla log-verosimiglianza:

$$\mathcal{L}(\Theta) = \sum_{i,j,\alpha} \left[A_{ij}^{(\alpha)} log \sum_{k,l} u_{ik} v_{jl} w_{kl}^{(\alpha)} - \sum_{k,l} u_{ik} v_{jl} w_{kl}^{(\alpha)} \right], \tag{2.6}$$

che porta alle stesse conclusioni di assumere che i Θ siano tutti uniformemente probabili a priori e usare il fatto che $\mathbb{P}(\Theta|G) \propto \mathbb{P}(G|\Theta)$. Si noti che il termine $A_{ij}^{(\alpha)}!$ al denominatore viene omesso data la sua dipendenza solo dai dati e non dai parametri.

La massimizzazione di (2.6) è computazionalmente difficile, perciò nell'articolo si propone un approccio variazionale e l'uso della disuguaglianza di Jensen per semplificare il calcolo. Si considera quindi un nuovo elemento $\rho_{ijkl}^{(\alpha)}$ che è la probabilità, per ogni coppia di gruppi k e l, che esista un arco tra i e j sulla base dell'appartenenza ai due gruppi dei due vertici rispettivamente. Dato il livello α, se una rete ha più archi tra gli stessi due nodi (è cioè un multigrafo), ogni arco avrà la sua probabilità ρ uguale per ognuno, dato che dipende solo dall'appartenenza dei nodi ai gruppi.

Si noti che la differenza tra un grafo semplice e un multigrafo si rispecchierebbe nelle successive formule tramite un'ulteriore sommatoria che permetterebbe di sommare le probabilità relative ad ogni arco. Questo risulta poi equivalente al considerare un $A_{ij}^{(\alpha)}$ di peso pari al numero di archi esistenti tra i e j, come risulta naturale dalla formulazione precedente. Per evitare un abuso di notazione, nelle successive formule non si considera questa ulteriore sommatoria, esplicitando il modello nel caso di archi singoli tra i nodi, ma tenendo a mente questa differenza (ovviamente nell'implementazione del modello sui dati, è stata utilizzata la modifica per un multigrafo).

Si può scrivere quindi

$$log \sum_{k,l} u_{ik} v_{jl} w_{kl}^{(\alpha)} = log \sum_{k,l} \rho_{ijkl}^{(\alpha)} \frac{u_{ik} v_{jl} w_{kl}^{(\alpha)}}{\rho_{ijkl}^{(\alpha)}}$$
$$\geq \sum_{k,l} \rho_{ijkl}^{(\alpha)} \, log \, \frac{u_{ik} v_{jl} w_{kl}^{(\alpha)}}{\rho_{ijkl}^{(\alpha)}} = \sum_{k,l} \rho_{ijkl}^{(\alpha)} \, log \, u_{ik} v_{jl} w_{kl}^{(\alpha)} - \sum_{k,l} \rho_{ijkl}^{(\alpha)} \, log \, \rho_{ijkl}^{(\alpha)},$$

che vale come uguaglianza quando

$$\rho_{ijkl}^{(\alpha)} = \frac{u_{ik} v_{jl} w_{kl}^{(\alpha)}}{\sum_{k',l'} u_{ik'} v_{jl'} w_{k'l'}^{(\alpha)}}.$$

A questo punto, massimizzare la (2.6) è equivalente a massimizzare la seguente

verosimiglianza

$$\mathcal{L}(\Theta, \rho) = \sum_{i,j,\alpha,k,l} \Bigg[A_{ij}^{(\alpha)} \Bigg(\rho_{ijkl}^{(\alpha)} \, log \, u_{ik} v_{jl} w_{kl}^{(\alpha)} - \rho_{ijkl}^{(\alpha)} \, log \, \rho_{ijkl}^{(\alpha)} \Bigg) - u_{ik} v_{jl} w_{kl}^{(\alpha)} \Bigg]. \quad (2.7)$$

Si può quindi massimizzare la 2.7 alternativamente aggiornando le stime di θ e ρ, tramite l'algoritmo EM (Dempster, Laird e Rubin 1977). Partendo da una soluzione iniziale per il valore dei parametri, il passo di *expectation* calcola il valore atteso della verosimiglianza condizionatamente alla stima dei parametri in quel passo, il passo di *maximization* calcola un aggiornamento dei valori stimati per l'insieme dei parametri in modo che massimizzino la verosimiglianza. I due step proseguono iterativamente fino al momento in cui è soddisfatta una regola d'arresto.

L'algoritmo EM non assicura la convergenza al massimo globale, quindi, per ovviare al possibile raggiungimento di un massimo locale e controllare la convergenza, si ottengono più stime corrispondenti a diverse inizializzazioni dei parametri, considerando come stima finale quella avente verosimiglianza (2.7) maggiore.

Si noti che le equazioni per il passo di massimizzazione sono ottenibili in forma chiusa tramite la derivata parziale di (2.7) rispetto ai vari parametri:

$$\begin{aligned} u_{ik} &= \frac{\sum_{j,\alpha} A_{ij}^{(\alpha)} \sum_l \rho_{ijkl}^{(\alpha)}}{\sum_l \big(\sum_j v_{jl} \big) \big(\sum_\alpha w_{kl}^{(\alpha)} \big)} \\ v_{jl} &= \frac{\sum_{i,\alpha} A_{ij}^{(\alpha)} \sum_k \rho_{ijkl}^{(\alpha)}}{\sum_k \big(\sum_i u_{ik} \big) \big(\sum_\alpha w_{kl}^{(\alpha)} \big)} \\ w_{kl}^{(\alpha)} &= \frac{\sum_{i,j} A_{ij}^{(\alpha)} \rho_{ijkl}^{(\alpha)}}{\big(\sum_i u_{ik} \big) \big(\sum_j v_{jl} \big)}. \end{aligned}$$

La questione fondamentale a questo punto è assegnare i nodi alle comunità secondo le informazioni che ci forniscono i vettori di appartenenza che si sono stimati. Fino ad ora non si è mai parlato dei valori che possono assumere i vettori u e v, ai quali non si era imposto nessun tipo di normalizzazione. È solo a questo punto che ci si preoccupa del fatto che essi rappresentano delle probabilità e quindi è necessario imporre che i loro valori sommino a uno tramite $\bar{u}_i = u_i / \sum_k u_{ik}$, in modo che risulti $\sum_k \bar{u}_{ik} = 1$. Lo stesso procedimento crea i vettori $\bar{v}_j$, mantenendo

le appartenenze in senso *incoming* e *outgoing* e quindi permettendo che un nodo appartenga a gruppi diversi in entrambi i sensi, oltre che con proporzioni diverse di gruppo in gruppo.

Riguardo a questo, se si volesse applicare una ulteriore restrizione per far appartenere ogni nodo a un solo gruppo (quindi avere gruppi disgiunti), basterebbe considerare il nodo appartenente al gruppo la quale entrata nel vettore è maggiore.

2.4.2 Scelta del numero di gruppi

Una delle criticità di questo tipo di modelli è la scelta del numero di gruppi, che non è direttamente ottenuta tramite la procedura di stima sopra descritta. Per questo motivo l'approccio comune consiste nello stimare più modelli in corrispondenza di diverse numerosità di gruppi e nel scegliere il migliore in termini di adattamento ai dati. Naturalmente l'obiettivo è la scelta del modello più accurato nel descrivere i dati a disposizione, mantenendo allo stesso tempo un obiettivo di interpretabilità e parsimonia.

Un metodo classico per la scelta del numero di gruppi è l'uso del *criterio di informazione di Bayes*, o semplicemente BIC (Schwarz 1978). Essendo in un contesto basato su un modello probabilistico, tramite la verosimiglianza e penalizzando per il numero di parametri del modello, è naturale il calcolo di questo criterio, la cui formula nel caso in esame è pari a

$$BIC = -2\log(\mathcal{L}^*) + (2NK + K^2L)\log(N_e)$$

dove con $log(\mathcal{L}^*)$ si denota il valore massimizzato della (2.6), con N il numero di nodi, con N_e il numero di archi (rappresentano le osservazioni), con K il numero di gruppi e con L il numero di livelli della rete.

Un altro metodo, usato nel lavoro di De Bacco et al. (2017), è l'uso dell'area sotto la curva ROC. Solitamente, la curva ROC viene usata in contesto supervisionato ed è una curva rappresentabile in due dimensioni, nel caso in cui si abbiano due gruppi e quindi si possa calcolare sensibilità e specificità del modello nel classificare le unità. Esistono comunque in letteratura delle generalizzazioni al caso multi-gruppo (Fawcett 2006).

Un modo per ridurre l'intera curva ROC a un singolo indice quantitativo di accuratezza, è il calcolo dell'area sotto la curva, chiamata AUC (Hanley e McNeil

1982), che essendo una porzione dell'area del quadrato unitario ha un valore che varia tra 0 e 1 (assume valore pari a 0.5 nel caso di assegnazione casuale ai gruppi). Hanley e McNeil (1982) mostrano l'equivalenza tra il calcolo dell'area sotto la curva ROC e la statistica di Wilcoxon, fornendo un'interpretazione utile per l'applicazione in questo contesto. Infatti, l'area AUC ha una proprietà statistica importante che esplicita l'equivalenza: l'AUC di un classificatore è equivalente alla probabilità che il classificatore assegni a un vero positivo estratto casualmente una probabilità più alta rispetto a un vero negativo estratto casualmente. Così definita, questa misura è chiaramente non parametrica, non dipende da assunzioni fatte a priori e si generalizza facilmente al caso con più classi, come mostrato da Hand e Till (2001). Si collega al caso in esame se si pensa a un arco esistente nella rete come a un vero positivo, e a un arco mancante come vero negativo.

Per testare l'accuratezza del modello e scegliere quindi il migliore (che avrà un determinato numero di gruppi) si sceglie di applicare una *cross-validation* suddividendo i dati in cinque parti. I dati in questo caso sono le entrate della matrice di adiacenza e si considerano tali sia le entrate positive che gli zeri; durante la stima del modello vengono usate l'80% delle entrate mentre il test del modello viene fatto sul restante 20%. Tramite la stima fatta nel cosiddetto *training set* si misura nel *test set* la capacità del modello di prevedere le entrate precedentemente nascoste, nel senso di dare una probabilità maggiore a un arco presente rispetto che a uno assente.

Si noti che essendo in un caso di archi multipli ci sono due aspetti da tenere in considerazione. Il primo riguarda il nascondere le entrate della matrice: essendo potenzialmente formate da più archi ognuna, va estratto casualmente anche il numero di archi da nascondere sul totale considerandoli quindi singolarmente in modo da testare l'abilità del modello per ogni arco. E secondo aspetto, anche per il calcolo dell'AUC partecipa la probabilità di ogni arco preso singolarmente.

2.4.3 Interdipendenza tra strati

Un aspetto che sicuramente merita una menzione a parte è lo studio dell'interdipendenza tra gli strati della rete. Quando si pensa a diverse tipologie di connessioni tra i nodi di una rete, ci si chiede anche come l'informazione sia ripartita tra questi

livelli ed eventualmente se sono tutti necessari, se c'è correlazione tra di essi e se è possibile quantificarne la forza.

De Bacco et al. (2017) hanno proposto un metodo per inferire su un tipo generale di interdipendenza, basata sull'idea che due strati sono collegati solo se la struttura di un livello fornisce informazione che migliora la stima dell'altro livello. Nello specifico, dopo aver stimato il modello, si fa previsione su un arco in uno strato con e senza informazione sull'altro strato e si valuta se l'aggiunta di informazione renda migliore la previsione, quantificando quindi se questi livelli sono interdipendenti.

Per la definizione di interdipendenza tra una coppia di livelli α e β, si consideri quanto detto in (2.4.2). In particolare, si procede facendo una previsione sul primo livello α, con il 20% delle entrate di $A^{(\alpha)}$ nascoste, ma fornendo in un primo momento solo le restanti entrate di $A^{(\alpha)}$, successivamente anche l'informazione fornita dall'intera matrice $A^{(\beta)}$. La differenza negli AUC tra questi due esperimenti determina quanto la conoscenza dello strato β aiuta nella previsione del livello α. In questo modo si definisce il *2-layer AUC* e in modo del tutto analogo si possono definire gli *l-layer AUC* con $l \in (1, ..., L)$ nel caso generale in cui sono presenti L livelli.

2.5 Altri possibili modelli

In un contesto di modelli probabilistici, gli *Stochastic blockmodels* sono un'ampia classe di modelli per grafi casuali dove si presuppone l'esistenza di una struttura di comunità latente e si cerca di inferire su di essa. Ne fanno parte diversi filoni di ricerca che cercano di generalizzare i modelli di base o comunque portano alcune modifiche all'originale. Per citare alcuni esempi, si possono trovare i modelli a variabile latente (Handcock, Raftery e Tantrum 2007), i modelli con spazio latente (Hoff, Raftery e Handcock 2002), i *degree corrected blockmodel* proposti dagli stessi autori come estensione del modello visto in (2.3) per comunità disgiunte (Ball, Karrer e Newman 2011; Karrer e Newman 2011) e infine i *mixed membership blockmodel* citati in precedenza.

La scelta del modello utilizzato con i dati raccolti è stata dettata dalla generalità dello stesso e dalla considerazione in un unico modello di tutti gli aspetti di interesse per il lavoro. Naturalmente altre scelte erano possibili anche tra i

modelli appena citati, valutando di volta in volta se le criticità fossero prese in considerazione.

Uno sviluppo sicuramente da prendere in considerazione per il modello scelto riguarda l'inserimento di covariate di nodo direttamente nel modello, che potrebbero migliorarne la stima con le informazioni a disposizione dalla raccolta dei dati. Altro aspetto menzionato in precedenza, è la possibilità di inserire eventuale informazione disponibile sui parametri tramite una o più distribuzioni a priori su di essi e quindi calcolando e usando la distribuzione a posteriori per l'inferenza.

2.6 Un algoritmo di ranking

Un aspetto di interesse differente rispetto ai modelli precedenti riguarda l'algoritmo usato per cercare di estrarre un ordinamento di tipo lineare tra i nodi della rete.

Il principio su cui si basa l'algoritmo viene detto *minimum violation ranking* e riguarda il calcolo del numero di inconsistenze per diversi possibili ordinamenti e lo scambio di unità in modo iterativo per ottenere, come si evince dal nome, il ranking con il minor numero di violazioni possibile (De Vries 1998).

In una rete direzionata G con matrice di adiacenza A, una violazione o inconsistenza è un arco (u, v) dove il rango (la posizione nel ranking) di v è migliore di quello di u. In merito a ciò, De Vries (1998) definisce una matrice quadrata dove per riga si inseriscono i "vincitori" e per colonna i "perdenti" e, tramite un algoritmo iterativo che di volta in volta scambia le posizioni di una coppia di celle, si conta il numero delle inconsistenze fino a minimizzarle e trovare un ordinamento lineare.

Una violazione è facilmente individuabile confrontando i valori nelle due celle A_{uv} e A_{vu} della matrice di adiacenza. Si denoti con π una permutazione degli N nodi nel grafo e con $\pi(u)$ l'indice dell'unità u nella specifica permutazione. Infine, con $\pi(A)$ si fa riferimento alla matrice di adiacenza riordinata sulla base di tale permutazione.

Un *minimum violation ranking* (MVR) è una permutazione π che induce il minor numero di archi nella rete verso posizioni superiori dell'ordinamento, le violazioni appunto, e che conseguentemente implica il maggior numero di archi che puntano verso il basso. Nell'approccio e nell'ambito applicativo considerato da

Clauset, Arbesman e Larremore (2015) e seguito in questa tesi, questo significa trovare un ordinamento degli atenei, per ognuno dei settori considerati, dove dall'ateneo di formazione a quello di lavoro il docente non si sposti verso un ateneo posizionato meglio nella classifica, ma peggio. Gli autori partono dal presupposto che se l'assunzione dei docenti seguisse una perfetta gerarchia sociale tra gli atenei, nessun docente verrebbe assunto in un'istituzione più prestigiosa rispetto a quella dove si è formato.

Il MVR si può ottenere massimizzando il peso degli archi che non comportano violazioni, calcolato come

$$S[\pi(A)] = \sum_{u,v} A_{uv} \times sign[\pi(v) - \pi(u)] \tag{2.8}$$

che sottrae il peso degli archi che violano l'ordinamento (contenuti nel triangolo inferiore della matrice di adiacenza permutata $\pi(A)$) dal peso degli archi che non violano l'ordinamento.

Il problema della versione deterministica dell'algoritmo teorizzato da De Vries (1998) è che, in reti complesse, può trovare diversi ordinamenti dei nodi che producono lo stesso minor numero di violazioni. Andrebbe quindi fatta una scelta arbitraria di uno dei ranking con uguale MVR.

Per questo motivo Clauset, Arbesman e Larremore (2015) propongono l'uso di un'ottimizzazione di tipo stocastico tramite *simulated annealing* (Kirkpatrick, Gelatt e Vecchi 1983), equivalente all'algoritmo MCMC con una specifica funzione di accettazione Metropolis-Hastings (Hastings 1970; Metropolis et al. 1953). Ad ogni passo l'algoritmo cerca la soluzione ottima scambiando la posizione di una coppia di nodi e accettando lo scambio qualora la proposta sia neutra o migliore; questo approccio viene detto con funzione di accettazione a temperatura zero, che implica la selezione di un parametro di controllo che faccia sì che cambiamenti verso l'alto del numero di violazioni vengano accettati con una probabilità che decresce a ogni iterazione convergendo a zero, mentre cambiamenti verso il basso od ordinamenti con lo stesso numero di violazioni vengano sempre accettati. In sostanza, invece di scegliere un determinato ordinamento tra quelli che esibiscono uno stesso MVR, si utilizza un approccio di tipo Monte Carlo e si itera l'algoritmo un numero elevato di volte con diverso ordinamento iniziale, scelto casualmente; una volta fatto questo, si considera per ogni unità la sua posizione media o mediana tra tutte

le iterazioni ma è importante notare che si dispone anche della sua distribuzione empirica.

Un appunto va fatto riguardo gli auto-archi. È chiaro dal procedimento appena spiegato che se un docente rimane nella stessa facoltà dove si è formato non c'è nessun confronto con un altro ateneo e quindi questo movimento non impatta sul ranking sotto l'equazione 2.8. Si considera quindi un auto-arco come un movimento il più piccolo possibile verso il basso.

Una critica possibile a questo metodo può essere fatta pensando che ogni violazione in questo metodo ha peso uguale: non vi è quindi differenza tra una situazione in cui vi è violazione tra due atenei distanti una posizione nel ranking oppure dieci posizioni. Una soluzione possibile è pesare il MVR per la differenza di rango tra le due unità, cioè $\pi(v) - \pi(u)$, in modo da considerare più grave una violazione se riguarda un salto di più posizioni.

Capitolo 3

Dati e analisi esplorative

3.1 I dataset utilizzati

I dati usati in questa tesi si compongono di due distinti dataset. Entrambi riguardano la carriera accademica di docenti attualmente strutturati in un ateneo italiano. Le analisi mostrate in questo capitolo sono state eseguite con il software statistico R (R Core Team 2012).

Nello specifico, il primo dataset si compone dei 660 docenti inquadrati in data 18/03/2016 nel macrosettore 13/D, statistica, e in particolare dei settori disciplinari da SECS-S/01 a SECS-S/05. Questi dati sono stati raccolti per una precedente tesi di laurea (Bush 2016), con l'obiettivo di seguire l'articolo di Clauset, Arbesman e Larremore (2015) e descrivere la struttura di questo settore valutando se esista una gerarchia tra atenei e se questa porti a disuguaglianze sociali tramite analisi descrittive e modelli logit.

Il secondo dataset riguarda l'ambito delle scienze fisiche ed è stato raccolto *ex novo* per questa tesi, per poter fare un confronto tra due settori disciplinari riprendendo la costruzione dell'articolo di riferimento. Si è deciso di scegliere un settore disciplinare vicino alla statistica in quanto materia scientifica portata anch'essa alla ricerca e quindi con propensione dei laureati a proseguire la carriera accademica, ma pensando anche a differenze sostanziali come per esempio che le scienze fisiche fanno parte delle così dette *hard sciences* mentre la Statistica è considerata una via di mezzo tra queste e le scienze umanistiche dal punto di vista dei temi di

ricerca, delle metodologie di studio e della collocazione dei lavori scientifici.

Una volta deciso di optare per l'ambito delle scienze fisiche, si è valutato che il settore è troppo ampio perché comprende tre macrosettori 02/A, 02/B e 02/C per un totale di 2166 docenti, in data 09/12/2016; è stato quindi necessario ridurre l'analisi al settore disciplinare FIS/01, fisica sperimentale, con un totale di 706 docenti, numerosità paragonabile all'altro dataset.

La nuova raccolta dati ha naturalmente rispecchiato quanto fatto nella precedente tesi per avere le stesse informazioni su entrambi i gruppi di docenti. Per la costruzione dei dataset la ricerca è iniziata nell'archivio CINECA, il consorzio italiano formato da settanta università, cinque enti nazionali di ricerca e il MIUR, che fornisce le informazioni di base riguardo tutti i docenti italiani, divisi per macrosettori e settori disciplinari. Le informazioni fornite riguardano nome e cognome dei docenti, genere, ruolo accademico o fascia, ateneo di lavoro attuale con relativo dipartimento al quale afferiscono i docenti e infine settori scientifico disciplinare e concorsuale.

Sono state raccolte manualmente per ogni professore informazioni relative all'intera carriera accademica, a partire dal percorso di istruzione (laurea, dottorato e post), passando attraverso tutti i ruoli assunti dal docente, segnando per ogni campo l'anno di inizio e fine, l'ateneo e il settore disciplinare o il dipartimento di afferenza (unica differenza tra i due dataset). Questi dati sono stati ricercati soprattutto nei curricula dei docenti, spesso inseriti nelle pagine personali presso i siti delle università, ma anche nelle valutazioni comparative pubbliche dei docenti svolte dagli atenei prima dell'assunzione degli stessi o in minima parte da altre fonti come per esempio i social network LinkedIn e ResearchGate.

La ricerca ha richiesto mediamente quindici minuti a docente in modo simile per i due settori e in linea con quanto dichiarato per il contesto americano.

3.1.1 Qualità del dato

Non è stato possibile reperire le informazioni complete per tutti i docenti. Per le specifiche esigenze di analisi di questo lavoro, le informazioni fondamentali riguardano l'istruzione (laurea e dottorato) e l'ateneo attuale di lavoro. Per quanto concerne quest'ultima informazione, la copertura nei dati è totale perché CINECA fornisce il dato sulla totalità dei docenti. Al contrario, per laurea e dottorato il tasso

di dati mancanti dipende soprattutto dalle informazioni inserite dal docente online. Essendo il dottorato stato introdotto in Italia solo nel 1980, i docenti laureati prima di quella data per la maggior parte non detengono un livello di studio superiore alla laurea (se non in alcuni casi una specializzazione). Per questo motivo verranno considerati come propriamente mancanti solamente i dati relativi alla laurea.

Nel dataset di Statistica, il tasso di dati mancanti per il campo ateneo di laurea è pari all'8% dopo una ricerca completa avvenuta da marzo a maggio del 2016. Nel dataset di Fisica (verrà da ora in avanti omesso il termine "sperimentale" per semplicità di lettura), il tasso di dati mancanti è maggiore e pari all'11% dopo una ricerca avvenuta da dicembre 2016 a febbraio 2017. Per chiarezza, questa seconda ricerca non si può definire completa in quanto non è stata approfondita la specifica ricerca dei docenti con dati mancanti per mancanza di tempo, ma si pensa che con ulteriori giorni a disposizione il tasso di dati mancanti avrebbe potuto raggiungere quello dei dati di statistica. Scopo di futuri lavori sarà innanzitutto quello di completare in modo il più esaustivo possibile questa ricerca.

Prima di proseguire con le analisi descrittive va analizzato in profondità un limite di questi dati e quindi delle analisi svolte in questo lavoro. Quella che l'Istat chiama *mobilità intellettuale* è la tendenza di sempre più dottori di ricerca di emigrare all'estero dopo aver concluso la formazione in Italia. Questo viene spiegato nel Rapporto annuale Istat 2015 dalle maggiori opportunità che i dottori di ricerca affermano di trovare all'estero, dove il lavoro è più qualificato e anche retribuito in modo migliore. In particolare, nel 2015 il 12.9% dei dottori di ricerca afferma di vivere abitualmente all'estero. Un aspetto limitante dei dati che si possono trovare in CINECA è proprio la mancanza di questa fetta di laureati o dottori di ricerca che hanno scelto di lavorare sì in ambito accademico ma all'estero, dei quali quindi viene persa l'informazione. Non si può perciò sapere il motivo per cui questi docenti abbiano scelto di andare all'estero e come il risultato finale sarebbe influenzato dalla presenza di questi dati.

Il problema inverso è molto meno rilevante: la netta maggioranza degli individui ha laurea e, se presente, dottorato entrambi svolti in Italia. Si hanno poi alcuni casi, con numerosità diverse tra Statistica e Fisica, di individui che hanno conseguito la laurea in Italia ma hanno scelto di iscriversi a un dottorato estero. Pochissimi, infine, sono i casi di docenti con formazione totalmente all'estero; allorché nulla la presenza di individui con laurea all'estero e dottorato di ricerca italiano. Si è quindi

scelto, vista la numerosità molto bassa di questo caso limite, di non considerare nelle applicazioni che verranno riportate in seguito questi individui con formazione estera per entrambi i titoli di studio, che invece saranno compresi nelle analisi descrittive iniziali riguardanti l'intero insieme dei docenti. Lo stesso vale per la rappresentazione grafica, che non avrà il nodo estero per una miglior leggibilità.

Tenendo a mente questi aspetti appena menzionati, si può definire l'obiettivo di questa tesi in modo più ristretto ragionando sulle informazioni a disposizione. Ciò che si può ottenere da un dataset con queste limitazioni è lo studio della mobilità italiana: si vuole cioè capire come si muovono gli individui che conseguono uno o entrambi i titoli in Italia e scelgono successivamente di rimanere in Italia a lavorare.

Si pensa comunque che quanto detto non diminuisca gli spunti di interesse dell'analisi, in quanto vengono messi in luce aspetti ancora poco studiati traendone idee rilevanti per la futura ricerca.

Nel seguito si mostreranno le analisi svolte sui dati prima da un punto di vista statico e solo successivamente come flusso tramite la creazione di una rete tra gli atenei italiani.

3.2 Analisi descrittive statiche

Tramite i dataset di cui si è parlato in (3.1) si possono ottenere analisi di tipo statico riguardo due aspetti separatamente. Si parla di analisi statiche in quanto, per ora, non viene considerato il movimento dei docenti tra gli atenei, ma si guarda una fotografia della situazione al momento della raccolta dei dati di entrambe le unità di interesse: atenei e docenti universitari.

3.2.1 Atenei

Riguardo gli atenei italiani, è stata fatta una ricerca nei siti istituzionali di alcune informazioni aggiuntive atte a descriverli: macro-regione geografica, dimensione, coordinate geografiche, anno di istituzione e tipologia.

Si è deciso di dividere in tre macro-regioni l'Italia, prendendo spunto dalla classica ripartizione di Istat: Nord, Centro e Sud.

- Il Nord comprende le regioni del Nord-Ovest (Liguria, Lombardia, Piemonte, Valle d'Aosta) e quelle del Nord-Est (Emilia-Romagna, Friuli-Venezia Giulia, Trentino-Alto Adige, Veneto);
- Il Centro comprende le regioni Lazio, Marche, Toscana ed Umbria;
- Il Sud comprende le regioni dell'Italia Meridionale (Abruzzo, Basilicata, Calabria, Campania, Molise, Puglia) e quelle dell'Italia insulare (Sardegna, Sicilia).

La dimensione è così classificata:

- Mega ateneo: le università con più di 40 000 iscritti totali;
- Grande ateneo: le università con numero di iscritti tra i 40 000 e i 20 000;
- Piccolo ateneo: le università che non raggiungono i 20 000 iscritti.

Infine, la tipologia di istituzione è rappresentata per la maggioranza dalle università statali, cioè istituite direttamente dallo Stato Italiano. Altre possibilità presenti in Italia sono le università private, sempre riconosciute da un decreto ministeriale del MIUR e promosse da enti pubblici o privati, le università telematiche e infine le scuole superiori universitarie.

Al momento in Italia sono presenti 97 atenei, tra cui: 68 atenei Statali (23 Nord, 20 Centro, 25 Sud), 18 atenei non Statali (9 Nord, 6 Centro, 3 Sud) e 11 atenei non statali telematici (1 Nord, 7 Centro, 3 Sud).

In generale, nei dataset sono presenti 76 diversi atenei, che si possono vedere in figura 3.1. Tra questi, gli atenei con almeno un docente di Statistica sono 69, mentre solo 55 per Fisica. Come si potrà capire meglio considerando i dati da un punto di vista dinamico, la motivazione plausibile di questo comportamento è il fatto che nel caso di statistica, essendo una materia di supporto per molte altre discipline, può essere presente un docente di questa materia anche negli atenei dalla più disparata offerta didattica. Per Fisica, invece, vale la regola generale che gli atenei dove si possono trovare docenti di fisica sono anche gli stessi dove questi si possono formare.

Una particolarità è data da alcune città dove sono presenti più atenei: i casi più rilevanti sono quelli Milano, Roma e Napoli che hanno rispettivamente sette, undici

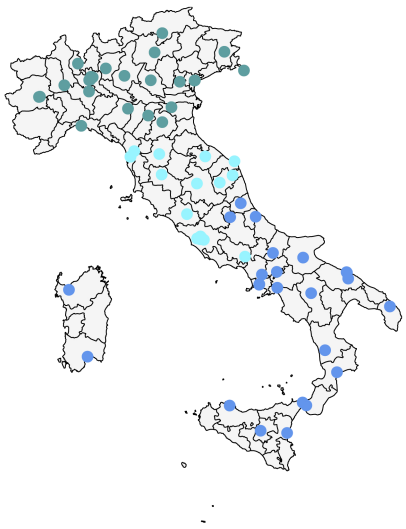


Figura 3.1: Mappa geografica degli atenei italiani

e quattro atenei. Casi minori riguardano Torino e Pisa con due e tre università. Gli atenei stranieri sono 20 per Statistica e 35 per Fisica, tre di questi in comune. Questi sono ripartiti tra atenei di laurea e dottorato, come si vedrà in dettaglio nel paragrafo 3.2.2, e sono tutti europei e americani.

Le ripartizioni degli atenei tramite le informazioni su macro-regione geografica e dimensione, per settore disciplinare, si possono vedere nelle tabelle 3.1 e 3.2. Le classi predominanti in entrambi gli ambiti sono la piccola dimensione, che rappresenta circa la metà degli atenei sia per Statistica che per Fisica, e il Nord, dove si trovano il 40% degli atenei di entrambe.

	Mega	Grande	Piccolo	Totali
Nord	4	9	14	27 (40%)
Centro	3	3	15	21 (30%)
Sud	4	7	10	21 (30%)
Totali	11 (16%)	19 (28%)	39 (56%)	69 (100%)

Tabella 3.1: Ripartizione geografica e tramite dimensione degli atenei nel dataset Statistica

	Mega	Grande	Piccolo	Totali
Nord	4	9	9	22 (40%)
Centro	3	3	8	14 (25%)
Sud	4	6	9	19 (35%)
Totali	11 (20%)	18(33%)	26(47%)	55 (100%)

Tabella 3.2: Ripartizione geografica e tramite dimensione degli atenei nel dataset Fisica

L'anno di istituzione può dividere gli atenei in tre gruppi: università antiche (fondate prima della Prima Guerra Mondiale), moderne (fondate fino agli anni Ottanta compresi) e nuove (dagli anni Novanta in poi). Tra gli atenei nuovi ci sono le università telematiche e va detto che queste offrono corsi soprattutto di tipo economico e nessuno offre una formazione in fisica (si veda tabella 3.3 per un riassunto generale).

Infine, per quanto riguarda la tipologia, mostrata in tabella 3.4, è predominante la presenza degli atenei statali ma in Fisica è molto più alta la proporzione sul totale. Si noti che le università telematiche sono state raggruppate nella categoria "privato". La categoria "altro" mette insieme gli atenei Kore di Enna e di Bolzano che non rientrano nelle università statali, ma sono promosse da due province italiane quindi non possono essere considerate come private.

Queste prime analisi servono, come detto, ad avere una fotografia generale della situazione degli atenei, che successivamente diventeranno l'unità di analisi della rete e, grazie ai flussi di individui che si sposteranno in essa, sarà possibile uno studio della struttura di collegamento ed eventualmente la ricerca di gruppi di atenei simili per quanto riguarda il pattern di reclutamento dei docenti.

	Generale	Statistica	Fisica
Antichi	29 (38%)	28 (41%)	25 (45%)
Moderni	26 (34%)	24 (35%)	19 (35%)
Nuovi	21 (28%)	17 (24%)	11 (20%)
Totali	76 (100%)	69 (100%)	55 (100%)

Tabella 3.3: Ripartizione tramite l'anno di istituzione degli atenei

	Generale	Statistica	Fisica
Statali	59 (77%)	54 (78%)	50 (90%)
Privati	12 (16%)	10 (15%)	3 (6%)
SSU	3 (4%)	3 (4%)	1 (2%)
Altro	2 (3%)	2 (3%)	1 (2%)
Totali	76 (100%)	69 (100%)	55 (100%)

Tabella 3.4: Ripartizione tramite la tipologia degli atenei

3.2.2 Docenti

In questo paragrafo l'analisi si focalizza sulla seconda unità di riferimento in questo lavoro: i docenti universitari. Nelle analisi di rete del capitolo successivo i docenti universitari rappresenteranno il peso degli archi che collegano i diversi nodi (atenei).

Nei dataset a disposizione sono presenti 660 docenti di statistica e 706 docenti di fisica sperimentale. Le prime analisi vengono riportate sulla totalità di questi dati, dei quali si hanno informazioni quali nome e cognome, genere, fascia e ateneo di riferimento al momento della ricerca dei dati.

Per quanto riguarda il ruolo assunto dai docenti nell'università in cui lavorano, vanno fatte alcune precisazioni. Innanzitutto, nei dati estratti da CINECA l'informazione relativa alla fascia del docente è inserita esattamente come definita dall'ordinamento e quindi sono presenti i ruoli sia come "confermati" che non per i ricercatori e gli associati ed esiste anche la figura del docente "straordinario" (ordinario non ancora confermato). Per semplicità si sono raccolte tutte queste figure nelle tre principali (ordinario, associato e ricercatore) e si vede in tabella 3.5 che la percentuale di docenti in ognuna è simile tra le due materie, con una predominanza di professori associati.

In secondo luogo, anche se la carriera accademica solitamente inizia con assegni di ricerca o borse di studio erogate da università o da altri enti o ancora con contratti di recente introduzione come ricercatore a tempo determinato, si è scelto di registrare queste tipologie di contratto come *pre-job* e considerare la carriera vera e propria solo dal primo contratto come strutturato del docente in modo di seguirlo nei tre passaggi fondamentali oltre che eventualmente negli spostamenti

	Statistica	Fisica
Ordinario	194 (30%)	175 (25%)
Associato	280 (42%)	349 (49%)
Ricercatore	186 (28%)	182 (26%)
Totali	660 (100%)	706 (100%)

Tabella 3.5: Ripartizione dei docenti tra i ruoli principali

tra atenei.

Un aspetto di diversità rilevante tra i due settori riguardo queste prime informazioni è il genere dei docenti universitari: nel totale dei docenti di fisica sperimentale, è solo il 19% la porzione di sesso femminile; la cosa interessante di questo dato è che rispecchia la stessa percentuale (esattamente il 20%) che si ritrova nell'area *02: Scienze Fisiche* nel suo complesso. Situazione completamente diversa si riscontra in statistica dove vi è un maggior bilanciamento per quanto riguarda il genere e si hanno il 46% di docenti femmine.

Questo argomento andrebbe sicuramente approfondito e andrebbe ricercata la motivazione di questo forte squilibrio. In particolare potrebbero esserci due cause: la prima a monte e cioè nel fatto che poche femmine scelgono di studiare fisica e optano per materie diverse; la seconda potrebbe trovarsi nella scelta delle università di assumere più laureati o dottori di ricerca maschi rispetto che femmine e quindi andrebbe studiata una possibile disuguaglianza sistematica nelle assunzioni. Per uno studio in merito sui dati americani effettuato dagli stessi autori dell'articolo di riferimento di questa tesi, si veda Way, Larremore e Clauset (2016). Riguardo questo secondo aspetto, sarebbe potenzialmente interessante, soprattutto per le assunzioni negli ultimi venti o trent'anni, utilizzare le valutazioni comparative per i passaggi di ruolo dei docenti. In questi documenti pubblici si trovano solitamente i curricula di tutti i docenti che hanno partecipato alla valutazione e viene fornito il giudizio dei commissari sia singolarmente che nel complesso, per arrivare alla scelta fatta. In questo modo si potrebbe avere una valutazione qualitativa del processo di selezione e si potrebbe inferire se sia presente una scelta sistematica dei maschi rispetto alle femmine. Si sottolinea che questa ricerca potrebbe colmare un altro limite delle analisi che si stanno facendo e cioè il fatto che guardando una fotografia della situazione italiana degli atenei e dei docenti non si riescono a dare

valutazioni sulle assunzioni, perché non si è registrata l'informazione su quali altri docenti erano papabili per quel determinato posto di lavoro.

Riguardo l'ateneo di afferenza dei docenti, analisi specifiche verranno mostrate in (3.3) tramite l'analisi descrittiva della rete. Va sottolineato però un aspetto rilevante in merito a questo e cioè che quando si misura la grandezza di un ateneo con i dati a disposizione, si sta descrivendo sì l'ateneo ma solo in relazione al numero di docenti di statistica o fisica al suo interno. Si sta usando questa misura perché si pensa che ad un maggior numero di iscritti corrisponda in generale un maggior numero di docenti, ma bisogna sottolineare che sono due concetti diversi.

Se si confronta la grandezza effettiva degli atenei rispetto alle due materie in esame, si possono notare differenze importanti in figura 3.2, che mostra il numero di docenti per entrambi i settori negli undici atenei mega posizionati in ordine di numero di iscritti. I casi più esemplificativi sono per esempio quello di Padova, che si posiziona sesta come grandezza degli atenei italiani ma in effetti nei due casi in esame ha una numerosità di docenti assunti che la posizionerebbe al secondo o terzo posto, e questo mostra l'importanza di questi due dipartimenti a Padova. Altra particolarità riguarda Milano, che ha pochi docenti di entrambe le materie rispetto a tutte le altre università classificate come mega.

Fino a questo momento si sono considerati i dati forniti dalla ricerca iniziale, che quindi erano presenti per la totalità degli individui. Nel seguito, si passa alle informazioni ricavate manualmente per ogni docente e va quindi aperta una parentesi per quel che riguarda i dati mancanti. Come si è detto, la differenza di dati mancanti per il campo laurea è di circa 3 punti percentuali (si passa dall'8% per Statistica all'11% per Fisica). Questi sono sicuramente dati mancanti nel senso proprio del termine, ma va fatto un discorso diverso per il dottorato.

In generale si può dire che le persone per le quali non è stata trovata la relativa informazione non hanno conseguito un dottorato di ricerca. Questa assunzione è stata fatta tenendo in considerazione l'esistenza dell'archivio delle tesi di dottorato di ricerca conseguite in Italia presso la Biblioteca Nazionale Centrale di Firenze (BNCF). Questo catalogo contiene tutte le suddette tesi ma possono esserci alcune discrepanze o anche la mancanza di alcune informazioni soprattutto per i primi cicli di dottorato, quindi non si possono escludere casi in cui anche per il dottorato di ricerca il dato sia mancante nel senso di non trovato, ma si è visto che sono numerosità basse.

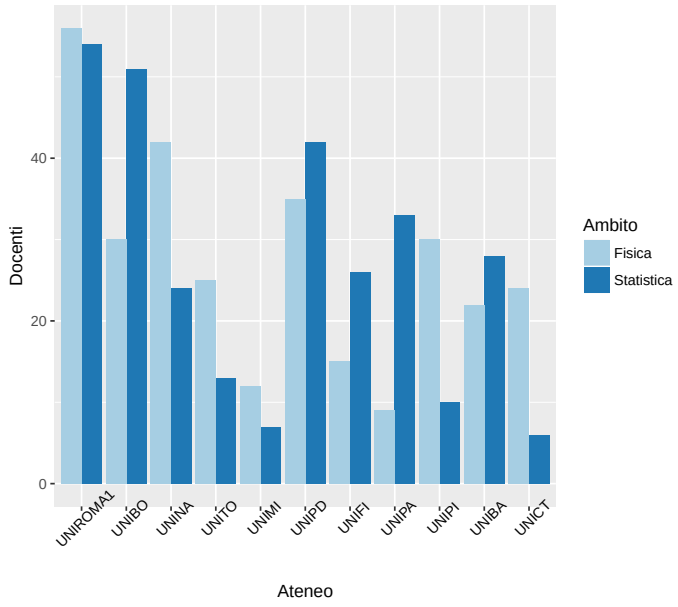


Figura 3.2: Numero dei docenti nei due ambiti disciplinari negli atenei mega italiani

Si noti che, per quanto detto, i docenti per i quali non si hanno contemporaneamente l'informazione su laurea e dottorato vengono eliminati prima delle analisi dal dataset e insieme a questi anche i pochi individui che hanno conseguito entrambi i titoli all'estero, data la decisione di non considerare il flusso dall'estero all'Italia. A questo proposito, come accennato in (3.1), si sono dovute prendere delle decisioni relative alla presenza di atenei esteri nella formazione dei docenti. Per Statistica, solo due persone hanno svolto la laurea all'estero e ventidue hanno conseguito il dottorato all'estero. Per Fisica la numerosità è maggiore: sette lauree e quaranta dottorati. In entrambe le situazioni, sono le stesse persone che hanno la laurea all'estero ad avere anche il dottorato all'estero e si è visto che sono cittadini non italiani che hanno deciso di venire successivamente a lavorare in Italia. Si è quindi deciso di eliminare questi nove individui considerandoli come con informazione mancante sia su laurea che dottorato.

Le considerazioni fatte riguardo i dati mancanti e i titoli conseguiti all'estero hanno portato a delle numerosità finali pari a 629 unità per Statistica e 630 unità

per Fisica. Dopo queste premesse, nei dati studiati si può dire che la mancanza del titolo di dottore di ricerca è pari al 26% per Statistica e al 33% per Fisica.

Si noti che nel seguito, quando verrà considerato il titolo maggiore conseguito dal docente, nel caso dei docenti con dottorato di ricerca all'estero verrà usata l'informazione riguardo la laurea conseguita in Italia, per quanto appena spiegato.

3.3 Analisi descrittive dinamiche tramite rete

Nello studio della rete creata a partire dai dati a disposizione non si è usata l'informazione sulla carriera, ma viene valutato solamente lo spostamento dall'ateneo di formazione al posto di lavoro attuale. Come spiegato, si hanno informazioni circa tutti i passaggi di ruolo dei docenti e anche dei trasferimenti tra diversi atenei. Uno studio possibile in questo caso è per esempio di tipo longitudinale per capire i movimenti nel tempo dei docenti.

Nel precedente paragrafo si è mostrato come sia possibile estrarre numerose informazioni di interesse anche esaminando i dati da un punto di vista statico, ma è chiaro che, vista la natura dei dati e lo scopo del lavoro, il considerare anche un aspetto dinamico possa arricchire quanto fatto.

Un modo per visualizzare il movimento dei docenti tra gli atenei è costruire una matrice di flusso, o matrice di adiacenza secondo la terminologia che si utilizza per le reti. Nel caso in esame questa è una matrice quadrata le quali righe e colonne corrispondono agli atenei italiani. La matrice contiene le informazioni riguardo il numero di docenti che si spostano tra ogni coppia di atenei. Riprendendo la notazione del capitolo precedente, A è una matrice $N \times N$ dove A_{ij} è la cella che incrocia l'ateneo i di formazione e l'ateneo j dell'attuale lavoro. Il numero presente nella cella è il flusso di docenti che si sono spostati da i a j. Una volta definita la matrice, il passaggio alla rete porta a definire gli atenei come nodi della rete e ogni docente come un arco tra due nodi.

Bisogna subito sottolineare che con i dati a disposizione si possono creare più matrici di questo tipo. Non considerando i campi relativi alla carriera del docente e quindi soffermandosi, come detto in precedenza, sulle informazioni riguardanti laurea, dottorato e lavoro attuale si possono costruire tre diverse matrici. La prima riassume l'istruzione in "istruzione maggiore" e quindi viene presa, se presente, l'informazione sul dottorato e, se non presente o se all'estero, quella sulla laurea.

La seconda e terza possono essere messe insieme in un tensore A di dimensione $N \times N \times 2$, tale che $A^{(1)}$ sia la matrice di flusso dei docenti dall'ateneo di laurea a quello di afferenza attuale e che $A^{(2)}$ rappresenti il flusso che parte dal dottorato di ricerca.

Una volta chiarito questo aspetto, risulta più intuitivo lavorare con una singola matrice rispetto che presentare tutte le analisi per ognuna delle componenti del tensore; per questo motivo, nel seguito di questo paragrafo si studierà il comportamento della matrice riguardante il passaggio dall'ateneo del maggior titolo di studio a quello del lavoro.

Questa tipologia di matrici è caratterizzata dal fatto di presentare usualmente una forte sparsità. Un grafico che riesce a cogliere in maniera esplicativa questo comportamento è la *mappa di calore* dove viene rappresentato tramite una sfumatura di colore il numero presente nella relativa cella (a partire dal bianco per le celle nulle). Nelle figure 3.3 e 3.4 le celle più visibili sono quelle della diagonale. Attraverso il grafico si possono inoltre individuare facilmente gli atenei dai quali e verso i quali ci sono dei flussi maggiori di docenti, che sono rappresentati rispettivamente da colonne o righe con un maggior numero di celle colorate. Il numero di celle contenenti un numero positivo nella matrice di adiacenza per il settore Statistica è pari al solo 5% sul totale, mentre in Fisica, pur avendo un numero di atenei minore che potrebbe indurre a pensare a una minore sparsità, il tasso di presenza è del 6%.

Alla luce di questi grafici un aspetto fondamentale che risalta riguarda la massiccia presenza di docenti che non si spostano e rimangono a lavorare, o tornano in alcuni casi, nello stesso ateneo dove si formano. Nel linguaggio proprio delle reti si tratta di una forte presenza di auto-archi. Anche se il 55% delle celle della diagonale della matrice di statistica contiene uno zero, il restante contiene il 45% del totale degli archi della rete. Per quanto riguarda Fisica, la percentuale di celle non nulle sulla diagonale è molto più alta e intorno al 64% ed esse contengono il 61% del totale degli archi della rete. I valori maggiori nella diagonale sono 45 *loops* in Statistica per La Sapienza di Roma, seguita da Bologna con un solo docente in meno, e 41 in Fisica sempre per Roma, seguita dai 34 del Politecnico di Milano. La particolarità di questi dati si rispecchierà anche nelle analisi che saranno presentate in seguito perché, nel caso di Fisica per esempio, è chiaro che la situazione che si prospetta è che solo il 40% dei dati permetterà lo studio degli spostamenti.

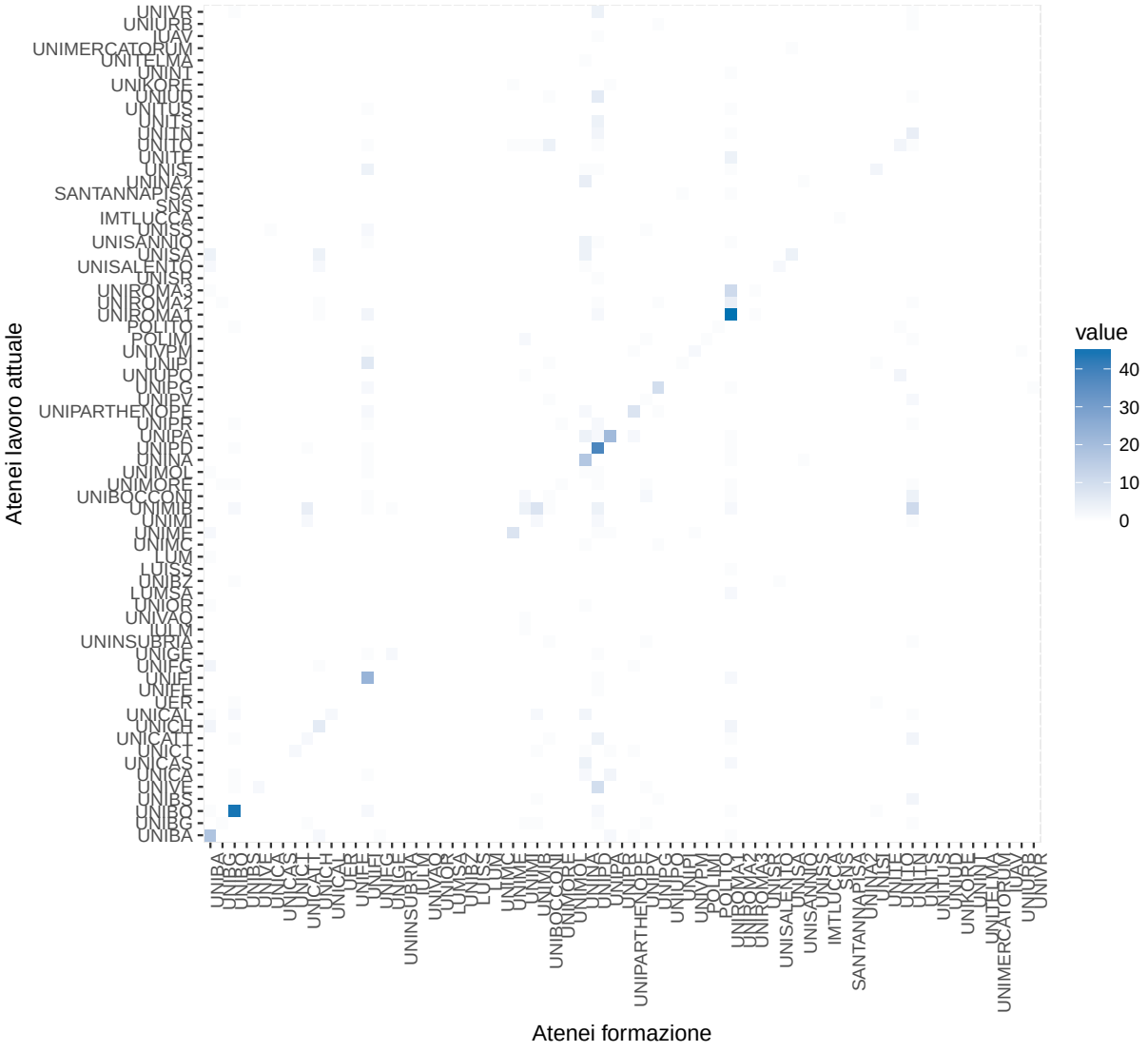


Figura 3.3: Matrice di adiacenza di Statistica tramite mappa di calore

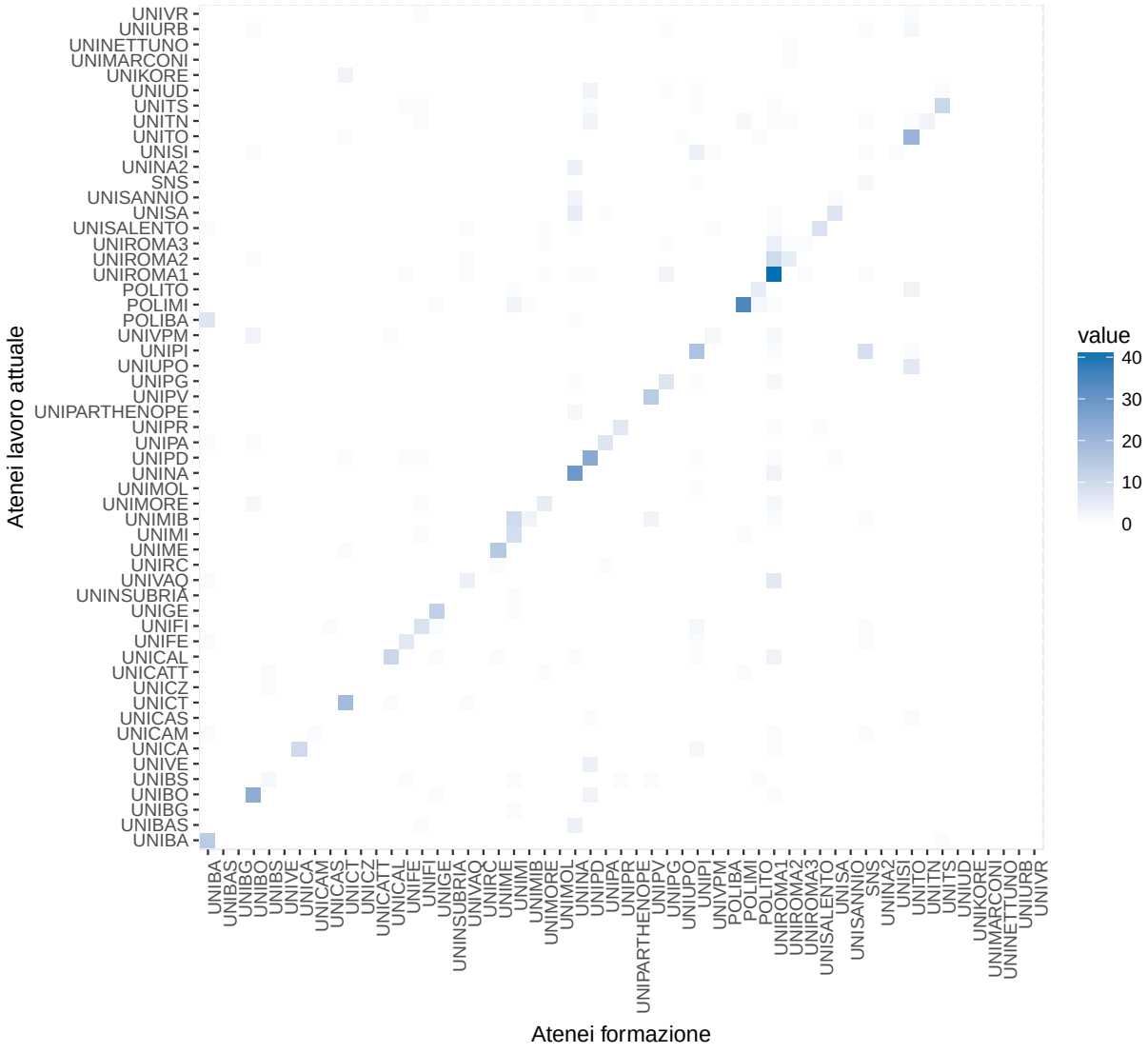


Figura 3.4: Matrice di adiacenza di Fisica tramite mappa di calore

Un altro aspetto descrittivo molto importante riguarda il grado del nodo, che si può calcolare in modo semplice considerando con quanti altri nodi esso ha almeno una relazione oppure pesando questo valore per il numero effettivo di archi a esso incidenti.

Partendo dal concetto più semplice, l'interesse è nel valutare la connettività degli atenei tramite il conteggio dei nodi con i quali condividono almeno un arco. Questo lo si può fare considerando il grado totale oppure l'*in-degree* e l'*out-degree*, in modo da farsi un'idea riguardo la specifica connotazione dell'informazione. Il grado in entrata di un ateneo dà informazioni sul numero di atenei da cui l'ateneo che si sta considerando assume laureati o dottori di ricerca mentre il grado in uscita stima il numero di atenei in cui esso riesce a posizionare i propri laureati. Si sottolinea che gli auto-archi sono stati eliminati da questo conteggio per parlare di connessioni tra atenei diversi, ma allo stesso modo si possono comprendere e contare la presenza di auto-archi aumentando di due l'ordine totale (è un arco che parte e arriva nello stesso nodo) e di uno le altre due tipologie.

Come si può vedere in tabella 3.6, dove sono ordinate le università con maggior grado nelle tre specificazioni per Statistica, la differenza sostanziale riguarda il grado in entrata.

Totale	UNIPD (29)	UNIROMA1 (28)	UNIFI (21)	UNITN (19)	UNINA (18)
In	UNIMIB (9)	UNITO (7)	UNIBG (6)	UNIMORE (6)	UNIROMA2 (6)
Out	UNIPD (25)	UNIROMA1 (24)	UNIFI (19)	UNITN (17)	UNINA (15)

Tabella 3.6: Classifica atenei più connessi per Statistica secondo il grado dei nodi della rete

Emblematico è il caso dell'università Milano Bicocca, che è un'università molto recente e forma poche persone connettendosi solo a sette università in uscita ma assume da un numero più elevato di atenei rispetto alle altre facoltà e quindi risulta al primo posto come grado in entrata ma non tra le prime cinque posizioni nel grado in uscita.

Altro aspetto interessante è notare che l'Università degli studi di Trento, anche se classificata come piccolo ateneo, è molto attiva nel formare futuri docenti di statistica ed entra nelle prime posizioni sia del grado in uscita sia del grado totale perché si connette con molte altre università. Allo stesso modo, La Sapienza di Roma è l'università più grande in Italia ma viene superata in due casi da Padova.

Totale	UNIROMA1 (28)	UNIPD (14)	UNIPI (14)	UNIFI (11)	UNINA (11)
In	UNIROMA1 (8)	UNITN (7)	UNIPD (6)	UNISAL (6)	UNIBS (5)
Out	UNIROMA1 (20)	UNIPI (11)	UNINA (10)	SNS (9)	UNIPD (8)

Tabella 3.7: Classifica atenei più connessi per Fisica secondo il grado dei nodi della rete

Ragionamenti analoghi si possono fare per Fisica (tabella 3.7), dove si nota subito il primato dell'Università La Sapienza di Roma in tutti i diversi gradi. Molto connessa in questo ambito è per esempio Pisa, che nel caso di Statistica non era presente nelle prime posizioni.

Un aspetto fondamentale che differenzia i due settori, accennato in precedenza, riguarda la presenza per Statistica di più università dove insegnano docenti iscritti al settore disciplinare ma dove nessun docente si è formato nello stesso. Trenta università (pari al 44%) hanno grado zero in uscita e si tratta delle facoltà dove la statistica è una materia di supporto ma non viene erogato nessun corso di laurea o dottorato di ambito inerente. In questi calcoli, si considerano quelle con grado zero contando anche gli auto-archi. Per Fisica, 18 università (circa il 30%) non presentano archi in uscita, cioè non formano docenti di fisica.

Contemporaneamente si può valutare non solo il numero di atenei con i quali il nodo è connesso, ma anche la quantità di docenti che si spostano tra quegli atenei. Ci può essere il caso di un ateneo molto connesso con un certo numero di altri perché manda un laureato in ognuno di questi, o al contrario un ateneo che forma moltissimi individui ma per qualche motivo questi vengono assunti tutti dallo stesso ateneo. Questo aspetto viene chiamato *forza* e si può intuitivamente pensare come la somma del peso delle connessioni tra i nodi, ragionando su collegamenti tramite un arco ognuno di peso pari al numero di docenti che passano per quella connessione. Vale lo stesso discorso per quanto riguarda gli auto-archi, che anche in questo caso vengono omessi visto che ci si è soffermati precedentemente.

Si sottolinea nuovamente il caso dell'università Milano Bicocca per Statistica, che è di nuovo molto esplicativo perché questo ateneo forma solo 16 persone, di cui 8 che rimangono a lavorare nella stessa università, ma ne assume 31 da altri atenei portandola al primo posto come *forza* in entrata. Succede lo stesso anche in Fisica perché questa università forma solo 4 individui (di cui tre rimangono a lavorarci) ma ne assume 15.

Totale	UNIPD (62)	UNIROMA1 (54)	UNIMIB (39)	UNITN (39)	UNINA (38)
In	UNIMIB (31)	UNIVE (12)	UNIROMA3 (12)	UNISA (12)	UNICATT (10)
Out	UNIPD (58)	UNIROMA1 (47)	UNINA (35)	UNITN (35)	UNIFI (34)

Tabella 3.8: Classifica atenei con il maggior numero di connessioni per Statistica secondo la forza dei nodi della rete

Totale	UNIROMA1 (54)	UNIPI (27)	UNINA (26)	UNIPD (23)	UNIMI (20)
In	UNIMIB (15)	UNIROMA2 (12)	UNIPI (11)	UNIROMA1 (10)	UNITN (10)
Out	UNIROMA1 (44)	UNINA (23)	UNIMI (18)	UNIPD (17)	SNS (17)

Tabella 3.9: Classifica atenei con il maggior numero di connessioni per Fisica secondo la forza dei nodi della rete

Dal confronto di (3.8) e (3.9) con le precedenti tabelle, si può notare che vi sono alcuni atenei che occupano la stessa posizione sia come grado che come forza, per esempio La Sapienza di Roma come grado totale e in uscita, sia in Statistica che in Fisica. Ci sono tuttavia dei casi particolari di università che "entrano in classifica" solo una volta tra tutti i casi visti, come per esempio l'Università degli Studi di Venezia che assume molti docenti di statistica che provengono da Padova e quindi risulta seconda come numero di docenti in entrata anche se tra le ultime come numero di connessioni con altri atenei (solo quattro). Si noti la presenza in entrambe le tabelle riferite a fisica della Scuola Normale Superiore che forma molti fisici che vengono assunti poi in numerose altre università.

Alla luce di quanto detto, si possono fare delle ulteriori considerazioni sul significato di *out-strength* e *in-strength*. Innanzitutto, la forza in uscita è sintomo della capacità che hanno gli altri atenei nell'assumere i docenti provenienti dal nodo che si sta considerando e quindi ipoteticamente non ha un limite superiore. Al contrario, la forza in entrata ha un chiaro limite che è pari al numero di posizioni aperte dall'ateneo e cioè al numero di docenti che può assumere, sia che essi provengano da altre università sia che vengano formati internamente. Il numero di docenti che è possibile assumere è quindi una misura della grandezza del dipartimento di statistica o fisica e/o della richiesta di docenti del settore in altri dipartimenti in quell'ateneo. Questo concetto sarà ripreso nel prossimo capitolo quando si parlerà di ranking e si vorrà valutare se la capacità di posizionare i propri laureati o dottori di ricerca è proporzionale o meno alla grandezza dell'ateneo.

A questo punto, è interessante vedere graficamente il grafo che si può costruire tramite la matrice di adiacenza definita a partire dal livello di istruzione maggiore, sia per Statistica che per Fisica. Si utilizzeranno le informazioni di cui si è parlato fino a questo punto.

La prima rappresentazione, che riprende figura 3.1 e si può vedere nella figura 3.5, mostra il posizionamento tramite coordinate geografiche degli atenei. Nel prossimo capitolo verrà spiegata nel dettaglio la costruzione di questa particolare tipologia di grafico.

In particolare, la grandezza dei nodi della rete è proporzionale al grado di ognuno nella sua specificazione totale, senza visualizzare i *loops* che si è scelto di omettere per una migliore rappresentazione. Il colore, poi, differenzia i nodi a seconda dell'area geografica a cui appartengono. Viene rappresentato un arco tra ogni coppia di nodi dove esiste almeno uno spostamento e lo spessore dell'arco è proporzionale al numero di individui che effettuano lo spostamento.

Per la seconda rappresentazione (figure 3.6 e 3.7) si aggiunge la forma dei nodi, che è circolare per gli atenei mega e quadrata per quelli grandi e piccoli.

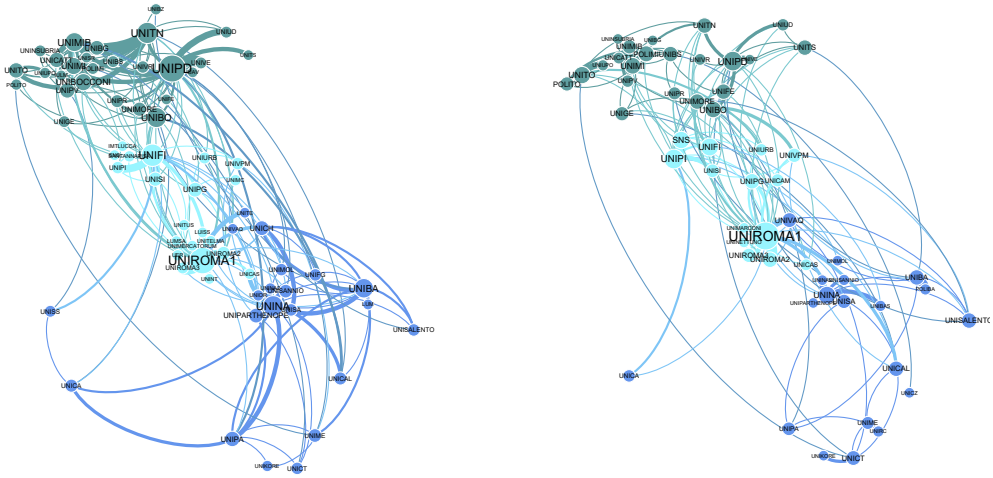
Il posizionamento dei nodi viene fatto tramite algoritmi ideati appositamente per fornire una resa grafica migliore della rete. Questi algoritmi calcolano il layout di un grafo usando solo informazioni contenute nella struttura stessa del grafo, senza affidarsi a conoscenze pregresse o esterne alla rete (Kobourov 2013). Uno di questi è stato ideato da Fruchterman e Reingold (1991) e utilizza i concetti, propri della fisica, di forza ed energia: si basa sull'idea che ci siano forze repulsive tra tutti i nodi ma anche forze attrattive tra nodi adiacenti. L'uso di questo algoritmo ha come obiettivo la rappresentazione della rete in modo graficamente ottimale ma anche intuitivo, dato che tende a tenere vicini i nodi più connessi tra loro e ad allontanare al contrario quelli non connessi.

Da (3.6) e (3.7) si possono notare diversi aspetti di contrasto, dai quali si possono estrarre diverse informazioni. Innanzitutto, una situazione particolare si verifica per Statistica, dove due atenei sono staccati dalla rete in quanto hanno un unico collegamento tra loro e nessun collegamento con la rete. Altri nodi, esterni rispetto al nucleo centrale nelle due reti, mostrano di essere meno connessi alla rete rispetto agli altri. Anche se in parte dipende dal numero di atenei e dai due atenei staccati che fanno sì che il grafico risulti più ristretto, è evidente che i nodi della rete di Statistica sono più vicini tra loro, a volte anche sovrapposti, a sottolineare

la forza maggiore dei collegamenti. Si nota poi una vicinanza più chiara dei nodi dello stesso colore, cioè nella stessa macro regione geografica, nella rete di Statistica e ciò mostra la presenza di più collegamenti interni a questi gruppi rispetto che verso gli altri atenei. Questo sembra meno chiaro nel grafo di Fisica, dove gli atenei sembrano disporsi in maniera più eterogenea al centro della rete.

La forma dei nodi ricorda quanto detto in precedenza sulla relazione tra dimensione degli atenei e dei singoli dipartimenti: il fatto di appartenere, per esempio, alla categoria mega non assicura di essere tra i nodi più grandi della rete. Roma è il nodo più grande in entrambe le reti, ma in statistica ci sono molti più nodi di grande dimensione e in generale sono più grandi rispetto ai corrispondenti in Fisica.

Si tengano presenti questi aspetti appena sottolineati per il confronto con i grafici che saranno mostrati nel capitolo successivo: per il colore dei gruppi non verranno usate informazioni a priori come l'area geografica, ma solo i risultati del modello stimato.



(a) Statistica **(b)** Fisica

Figura 3.5: Reti con nodi disposti tramite coordinate geografiche

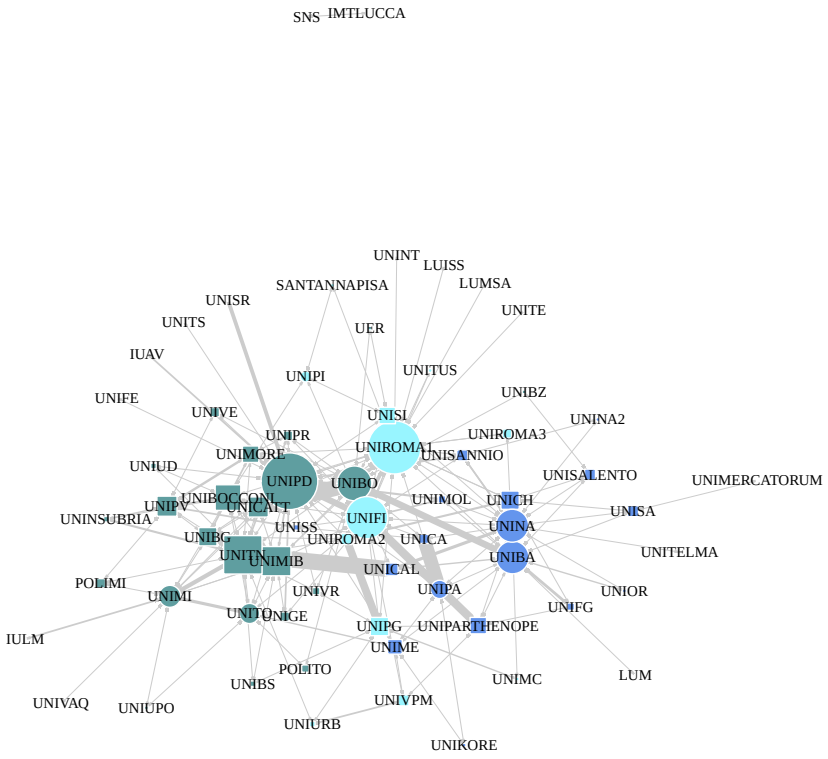


Figura 3.6: Rete di Statistica con nodi disposti tramite algoritmo di Fruchterman-Reingold

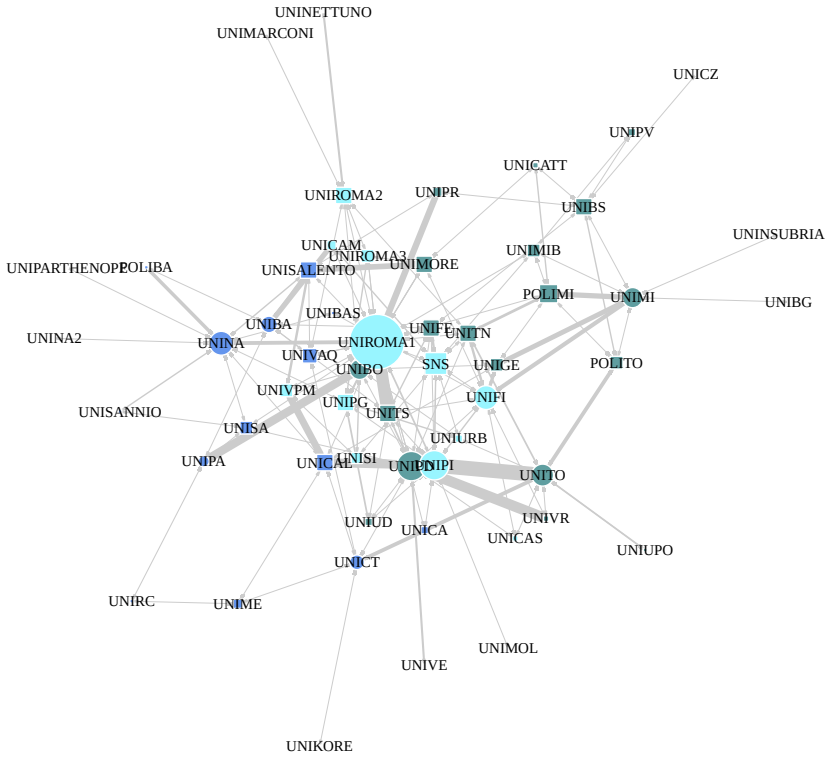


Figura 3.7: Rete di Fisica con nodi disposti tramite algoritmo di Fruchterman-Reingold

Capitolo 4

Un'applicazione alla mobilità dei docenti italiani

4.1 Community detection

Uno degli obiettivi principali nello studio delle reti è la ricerca di una descrizione delle stesse tramite la divisione dei nodi in gruppi a seconda della struttura di connessioni esistente. L'interesse nel trovare dei gruppi di nodi è il riflesso del naturale dividersi dei nodi di molte reti reali. Come visto, sono numerosi i modelli utilizzabili in questo contesto; allo stesso tempo, non tutti i modelli sono adeguati a ogni dataset in quanto non vengono sempre colte tutte le particolarità presenti.

Lo *stochastic blockmodel* multilivello per comunità miste è stato scelto per lo studio di questi dati perché riesce a cogliere in modo adeguato diversi aspetti di interesse:

- la presenza di auto-archi, predominante nei dati in esame;
- il peso degli archi tra i nodi, o ugualmente la presenza di più di un arco tra la stessa coppia di nodi;
- la direzionalità degli archi, per considerare in modo diverso se a un nodo arrivano o da esso partono gli archi;
- la possibilità di individuare gruppi tra i nodi della rete, permettendo anche l'appartenenza degli stessi a comunità miste;

- la possibilità di modellare più di uno strato nella rete, operando diverse scelte della matrice di adiacenza e sfruttando più informazione.

Per questi motivi, nel seguito sarà presentata l'applicazione del modello di De Bacco et al. (2017) sui dataset di Statistica e Fisica. Per l'applicazione del modello sui dati si è usato il software Python (Rossum 1995) mentre per la rappresentazione dei grafi si è usato Gephi (Bastian, Heymann e Jacomy 2009). Per entrambi i settori sarà stimato il modello sia con il solo livello costruito dall'istruzione maggiore, sia con i due livelli separati per laurea e dottorato. Oltre a questo, verrà valutata l'interdipendenza degli strati per capire se l'utilizzo di due livelli al posto di uno sia informativo.

A parte l'aspetto statistico e modellistico dell'applicazione di questo modello, l'interesse fondamentale è di tipo interpretativo. La ricerca di un ranking tra le università, come si vedrà in seguito, ha come motivazioni principali da un lato la possibilità per lo Stato di erogare fondi in modo proporzionale al prestigio e alla capacità degli atenei e dall'altro la possibilità per i futuri fruitori di un determinato corso di studi di valutare l'ateneo migliore. Invece, la capacità di raggruppare le università simili da un punto di vista contemporaneamente di capacità di formazione degli studenti e di modalità di assunzione post-formazione ha un interesse soprattutto descrittivo e di confronto fra diversi settori e/o contesti culturali. Una volta scelto tramite il modello il numero di gruppi adatto per quel determinato insieme di dati, le comunità identificate saranno valutate e interpretate.

4.1.1 Statistica

Come spiegato in precedenza nel capitolo teorico, ci sono molteplici aspetti da considerare nel commento dei risultati del modello applicato ai dati. Per ognuno dei settori disciplinari, si possono studiare diverse combinazioni dei seguenti aspetti:

- la costruzione tramite uno o due livelli della rete;
- il numero di gruppi scelto dall'algoritmo e/o dal criterio;
- le due diverse comunità stimate dal modello: la comunità in entrata e in uscita dei nodi e la comunità normalizzata formata dall'unione delle due;
- la struttura della rete stimata.

Partendo dalla scelta del numero di gruppi, per Statistica i due metodi utilizzati sono concordi nella scelta di due comunità di nodi sia con la costruzione a uno che a due livelli. Per il dettaglio dei risultati di entrambi i metodi di scelta si vedano tabella 4.1 e figura 4.1, ricordando che il BIC va minimizzato mentre il valore dell'AUC, che rappresenta l'accuratezza del modello, al contrario va massimizzato.

	$K = 2$	$K = 3$	$K = 4$
$L = 1$	3 400.36	4 239.87	5 134.27
$L = 2$	4 882.39	5 827.30	6 804.02

Tabella 4.1: Valori del BIC per $K = 2, 3, 4$ per Statistica con uno o due livelli (L)

Per quanto riguarda l'AUC, la scelta non è così netta se si considerano le bande di variabilità per ogni valore. Per questo motivo e per un interessante confronto interpretativo, verranno presentati entrambi i risultati con $K = 2$ e $K = 3$ gruppi. Si noti che con questo metodo è difficile ottenere una risposta chiara per uno specifico numero di gruppi in quanto si calcolano le bande di variabilità per il valore dell'AUC ed esse spesso si intersecano. Ciò comunque porta più flessibilità al metodo, portando a un'interpretazione più ampia.

Partendo dalla costruzione più semplice con il livello di istruzione maggiore e con due gruppi, è interessante il confronto con alcune delle analisi descrittive viste precedentemente. Quando si è parlato di *out-degree* e *in-degree*, si è visto come ci fossero dei casi di università con uno di questi due gradi pari a zero, cioè università che rispettivamente non formano statistici o non ne assumono. Risulta intuitivo

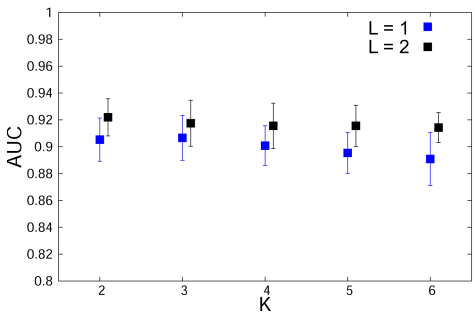
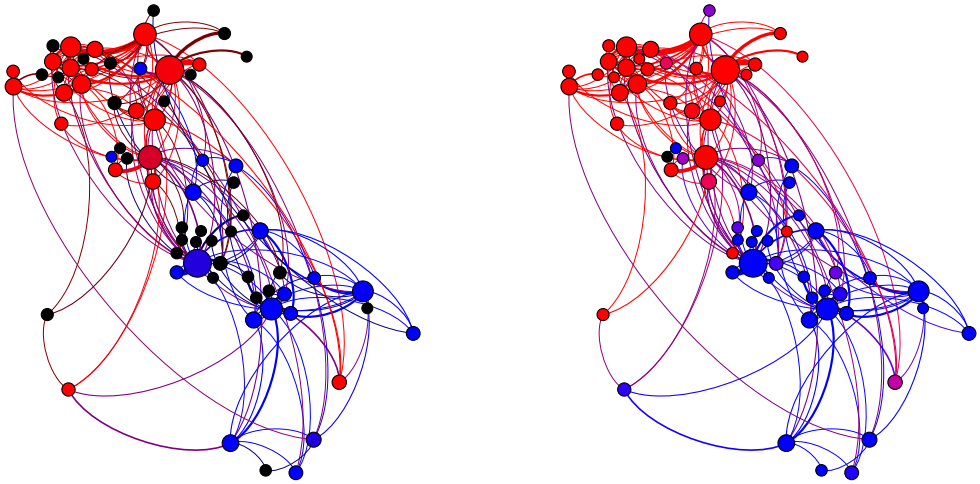


Figura 4.1: Valori dell'AUC per $K = 2, ..., 6$ per Statistica con uno o due livelli (L)

pensare che il modello, dovendo assegnare una comunità a questi particolari nodi, non riesca a farlo per entrambe le tipologie di comunità. Infatti, nelle figure 4.2a e 4.2b, dove i nodi sono colorati tramite i colori rosso e blu per mostrare i due gruppi stimati, si vedono alcuni nodi neri che corrispondono esattamente a quei nodi con valore pari a zero del corrispondente grado. Come già sottolineato, il problema maggiore si ha nel caso della comunità in uscita perché molte università non formano individui in questo ambito disciplinare, mentre il caso inverso riguarda solo la Scuola Normale Superiore.

Oltre a questo aspetto, si può notare la presenza di alcuni nodi che cambiano nettamente colore guardando l'una o l'altra tipologia di gruppo. Anche se la maggioranza dei nodi mantiene lo stesso colore nella stima delle due comunità, se questo non succede si tratta di atenei il cui comportamento è diverso tra la formazione e l'assunzione di docenti.



(a) Comunità in uscita **(b)** Comunità in entrata

Figura 4.2: Statistica: rete con un livello tramite le comunità in entrata e in uscita, $K = 2$ gruppi

Riguardo la costruzione del grafico, da qui in avanti è stata utilizzata l'informazione sulle coordinate geografiche per posizionare i nodi. Per le città dove

sono presenti più atenei, i nodi sono stati sparsi leggermente in modo che nessuno risultasse sovrapposto.

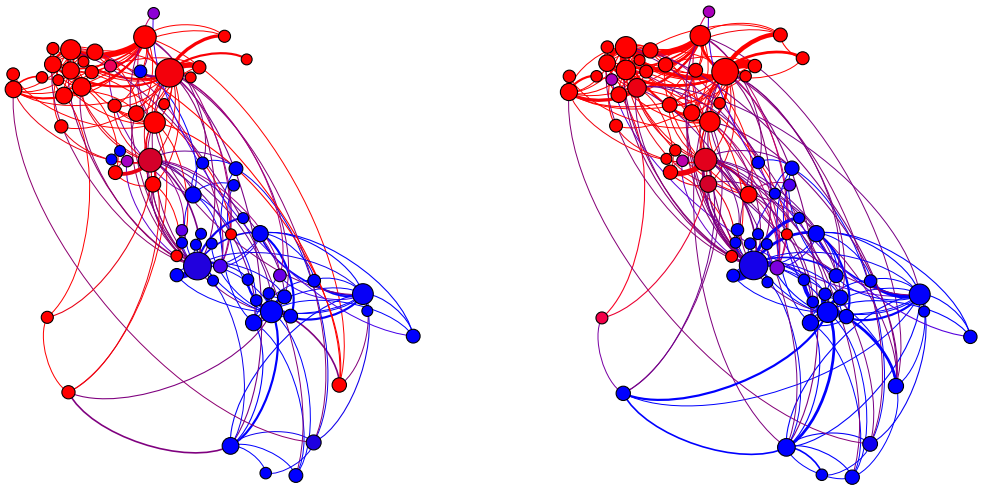
Per favorire l'interpretazione delle comunità viene operata una successiva semplificazione: si crea un'appartenenza normalizzata per i nodi, che racchiude l'informazione di entrambe le comunità ed è facilmente rappresentabile da un'unica figura (si veda (4.3a)). Si noti che i colori non sono per tutti i nodi ben definiti in rosso o blu ma alcuni hanno una certa sfumatura data dal mescolamento dei due colori principali. Come si è detto, il modello permette l'assegnazione di comunità miste per ognuno dei nodi e questo si rispecchia in alcuni nodi che, in questo caso, possono trovarsi tra i due gruppi con diversa proporzione di appartenenza.

L'obiettivo del lavoro è individuare dei gruppi di atenei simili tra loro sotto il punto di vista della formazione e dell'assunzione di docenti in quel determinato settore disciplinare. L'interpretazione che se ne dà tramite la visualizzazione è successiva alla stima del modello, che appunto stima le appartenenze ai gruppi senza nessuna informazione oltre la struttura della rete. In questo caso l'interpretazione geografica è la più intuitiva da intraprendere: si vede una chiara separazione, meno una decina di atenei, tra il nord Italia insieme alle regioni Toscana e Sardegna e il centro insieme al sud Italia.

Se si confrontano le attribuzioni dei nodi ai due gruppi operate dal modello in presenza di uno o due livelli (rispettivamente figure 4.3a e 4.3b), risulta chiaro che la presenza di più informazione tramite i due livelli presi singolarmente porti a una più netta divisione nei due gruppi geografici in quanto, nel caso dell'istruzione maggiore, il modello usa solo l'informazione del dottorato, se presente, mentre con due livelli l'informazione su entrambi i titoli viene usata in modo completo.

Le situazioni particolari di nodi singoli che si trovavano nella zona geografica del gruppo opposto non sono più presenti, restando comunque alcuni nodi con appartenenza mista ai due gruppi. Usare i due strati separati porta visivamente a un aumento degli archi, perché vengono rappresentati tutti quelli disponibili che partono dall'ateneo di conseguimento del titolo.

Questa interpretazione geografica è molto intuitiva perché si può pensare a gruppi di università molto connesse tra loro nel senso che si scambiano molti individui che si spostano vicino geograficamente, oltre alla grossa quota di individui che rimangono a lavorare nell'università dove hanno studiato, come si è visto dalle analisi descrittive.



(a) Rete con un livello **(b)** Rete con due livelli

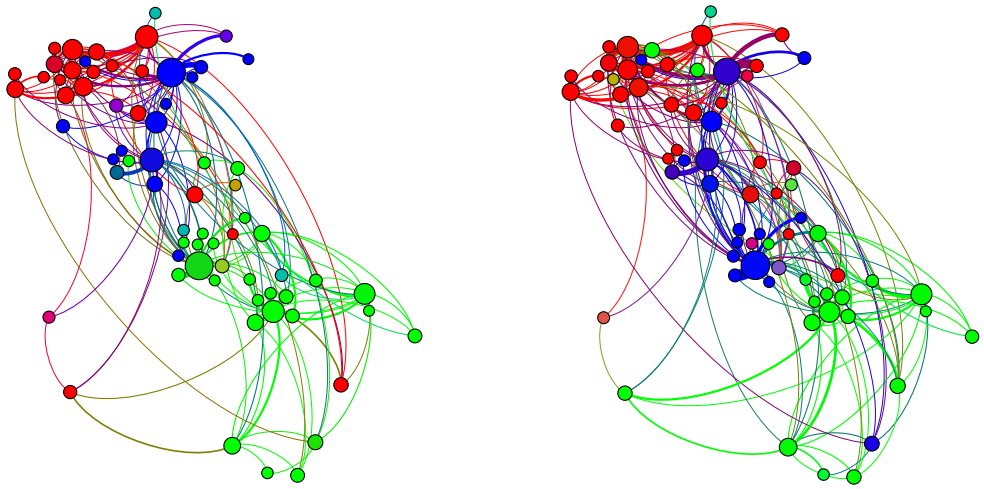
Figura 4.3: Statistica: rete con uno e due livelli tramite la struttura di comunità normalizzata, $K = 2$ gruppi

Passando alla stima del modello con tre gruppi, il primo aspetto di interesse è la comprensione di come si forma il terzo gruppo. Come si può vedere in figura 4.4a, soffermandosi inizialmente nel caso di un livello, il nuovo gruppo si crea dividendo il gruppo che corrispondeva al Nord Italia in una parte a nord-ovest che comprende anche la Sardegna e alcuni nodi geograficamente sparsi, e in una seconda parte che comprende le università di Padova, Bologna e Firenze e gli atenei maggiormente connessi a queste tre maggiori.

Anche se da (4.4a) la suddivisione in gruppi poteva sembrare gerarchica, dal confronto con (4.4b) si evince che non è così e infatti il modello dà una stima dei gruppi ogni volta in modo indipendente.

È interessante discutere dell'appartenenza a due diversi gruppi dell'Università Roma La Sapienza. L'aggiunta dell'informazione riguardante i due strati ha portato a un netto cambio di gruppo di questa università e delle più piccole università nel suo intorno geografico maggiormente legate a essa. Nel passaggio tra le due diverse strutture, si sono aggiunti degli archi tra La Sapienza e Bologna e Padova, due

atenei importanti nella rete relativamente alla loro grandezza. Questo ha portato La Sapienza a legarsi con il gruppo blu, che a questo punto mette insieme alcune delle università più importanti per questa rete, che evidentemente sono molto connesse tra loro e si scambiano molti individui. Si ha inoltre un effetto cascata di tutte le università che dipendono strettamente da La Sapienza, che insieme a questa confluiscono nel nuovo gruppo.



(a) Rete con un livello **(b)** Rete con due livelli

Figura 4.4: Statistica: rete con uno e due livelli tramite la struttura di comunità normalizzata, $K = 3$ gruppi

Oltre ai vettori di appartenenza ai gruppi, il modello stima la matrice strutturale della rete dato il numero di gruppi scelto. In questa rete la struttura è totalmente *assortativa* per ogni numero di gruppi, in quanto la matrice stimata presenta valori positivi lungo la diagonale e zero fuori. Questo si verifica sia nel caso di un livello, sia nel caso di due livelli dove viene stimata una matrice strutturale per ogni strato. Ciò significa una predominanza delle connessioni interne ai gruppi rispetto che a quelle tra nodi appartenenti a comunità diverse. Ad esempio, nei grafici con $K = 2$ queste ultime sono rappresentate tramite archi di colore viola perché mescolano il blu e il rosso dei nodi a cui sono connessi.

Aggiungendo il secondo livello nell'analisi si ha il passaggio da una configurazione semplice a una più complessa e ciò porta intuitivamente all'aggiunta di ulteriore informazione; questo apporto informativo può essere misurato tramite lo studio dell'interdipendenza tra strati. Come spiegato nel capitolo precedente, tramite l'AUC si può misurare quanto la conoscenza di uno dei due strati possa aiutare nella stima del secondo, e viceversa. Quanto si può evincere da figura 4.5 è che l'aggiunta del secondo livello aiuti nella previsione del primo e quindi ha senso l'uso del livello laurea e del livello dottorato in modo separato ma all'interno dello stesso modello. L'utilizzo di uno solo di questi livelli porterebbe a un'inferenza peggiore e ciò vale per ogni scelta del numero di gruppi.

Per la lettura della figura, si consideri inizialmente l'AUC per laurea e dottorato presi singolarmente (denotato con $L = 1$), il quale viene calcolato nascondendo il 20% delle entrate della relativa matrice e usando il restante 80% per la previsione. Passando invece alla notazione $L = 2$, per ognuno dei titoli si considera di nuovo il calcolo nascondendo il 20% delle entrate della stessa matrice, ma per la previsione, oltre alla restante porzione dell'80%, viene usata anche l'intera informazione relativa all'altro titolo di studio. In questo caso si può commentare, oltre quanto detto, che l'aggiunta dell'informazione sulla laurea aiuta di più rispetto a quella del dottorato e ciò si evince dall'ampiezza maggiore della differenza tra $L = 1$ e $L = 2$ Dottorato.

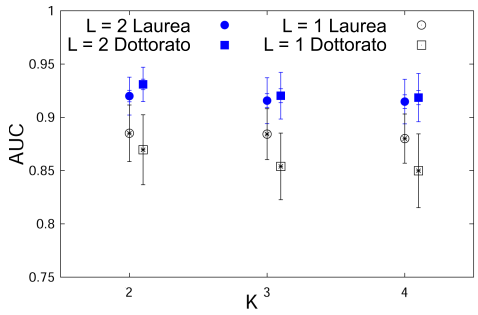


Figura 4.5: Valori dell'AUC per $K = 2, 3, 4$ per valutare l'interdipendenza degli strati in Statistica

Si noti che in questo lavoro, quando si è usato il singolo livello con l'istruzione maggiore, si è cercato di riassumere in un unico strato l'informazione su entrambi i titoli di studio. Il concetto di interdipendenza tra strati aiuta a capire se vale la

pena usare i due livelli laurea e dottorato insieme o separatamente ma non fornisce informazioni sul livello singolo così come costruito, in quanto l'informazione fornita è diversa.

4.1.2 Fisica

Seguendo gli stessi passi del paragrafo 4.1.1, va innanzitutto scelto il numero di gruppi migliore per questo settore disciplinare. In questo caso non c'è concordanza nei due metodi usati nella scelta dei gruppi: il BIC (tabella 4.2) opta ancora una volta per la presenza di due gruppi nei dati, penalizzando molto la crescita del numero di parametri con l'aumento del numero di gruppi; l'AUC (figura 4.6), che invece valuta la capacità previsiva del modello senza considerare nessun tipo di penalizzazione, porta a una scelta con K più alto, tra i 4 o 5 gruppi. Si noti che anche e soprattutto in questo caso il secondo metodo non fornisce una risposta molto chiara ed è difficile operare una scelta in questa situazione.

Anche se, da un lato, una migliore capacità del modello porta a una divisione più adeguata delle università in gruppi, dall'altro, situazioni con un numero elevato di gruppi sono difficilmente interpretabili. Dato il numero abbastanza ridotto di unità, al crescere del numero di gruppi si dovrebbe interpretare l'unione di pochi atenei fornendo poche informazioni realmente interessanti. Inoltre, bisogna sottolineare che il guadagno in accuratezza che si verifica nel passaggio da un numero di gruppi al successivo non è tale da giustificare una scelta più complicata dal punto di vista interpretativo.

	$K = 2$	$K = 3$	$K = 4$
$L = 1$	2 997.84	3 691.10	4 387.25
$L = 2$	4 156.56	4 922.33	5 719.75

Tabella 4.2: Valori del BIC per $K = 2, 3, 4$ per Fisica con uno o due livelli (L)

Per questo lavoro, si è scelto di valutare diverse numerosità di gruppo per avere un'idea più chiara dei cambiamenti tra una scelta e l'altra: in particolare si mostrano le situazioni con $K = 2, 3, 4$ gruppi.

Questa situazione meno netta per Fisica può essere causata in primo luogo dall'inferiore quantità di informazione sulle connessioni tra università diverse for-

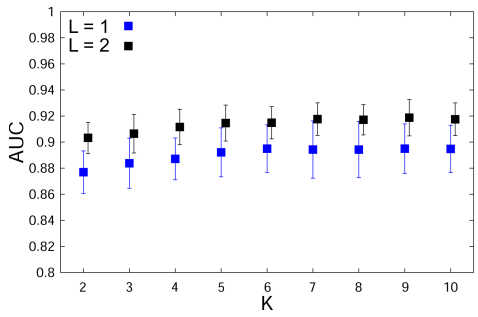


Figura 4.6: Valori dell'AUC per $K = 2, ..., 10$ per Fisica con uno o due livelli (L)

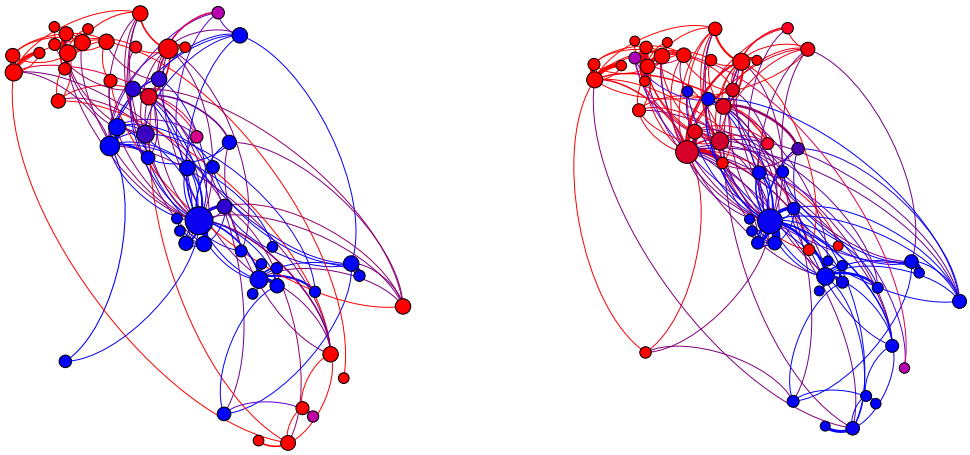
nita dalla rete: come spiegato nelle analisi descrittive, anche se le numerosità di partenza nei due settori disciplinari sono sostanzialmente uguali, in Fisica circa il 60% degli archi esistono sotto forma di *loops* e quindi meno della metà consistono in connessioni tra diverse università. In secondo luogo, una maggior mobilità degli individui in questo settore disciplinare potrebbe portare a una situazione meno netta. Riguardo questo aspetto, nel Rapporto annuale Istat 2015 viene sottolineato che l'area delle Scienze Fisiche è tra le più attive riguardo la mobilità intellettuale e questo indica che in generale è presente più mobilità verso l'estero rispetto a Statistica e può essere questo uno dei motivi per cui si creano in modo meno netto gruppi di università.

Iniziando dalla rete con due gruppi, la numerosità scelta dal criterio BIC, si nota una netta differenza nella composizione dei gruppi passando da uno a due livelli. Mentre in Statistica la separazione dei livelli portava a una conferma della configurazione con due gruppi, in questo caso l'interpretazione differisce.

Considerando il livello di istruzione maggiore (figura 4.7a), si crea un gruppo centrale che comprende le università dall'Emilia alla Campania e Puglia, insieme a Cagliari, con l'esclusione delle università più estreme a sud e parte della Sicilia, che sono raggruppate con il Nord Italia.

Se invece si considerano entrambi i livelli (figura 4.7b) la suddivisione si avvicina molto a quella geografica vista per Statistica in (4.3). A parte alcuni nodi nella zona geografica opposta al proprio gruppo, sembra ripresentarsi una divisione di tipo geografico tra Nord e Sud. Il "confine" si trova ancora una volta al di sotto della Toscana e virtualmente a metà delle Marche. Si noti il caso particolare di Cagliari che cambia gruppo nel passaggio da uno a due livelli, ma in modo opposto

a quanto succede nel settore Statistica.



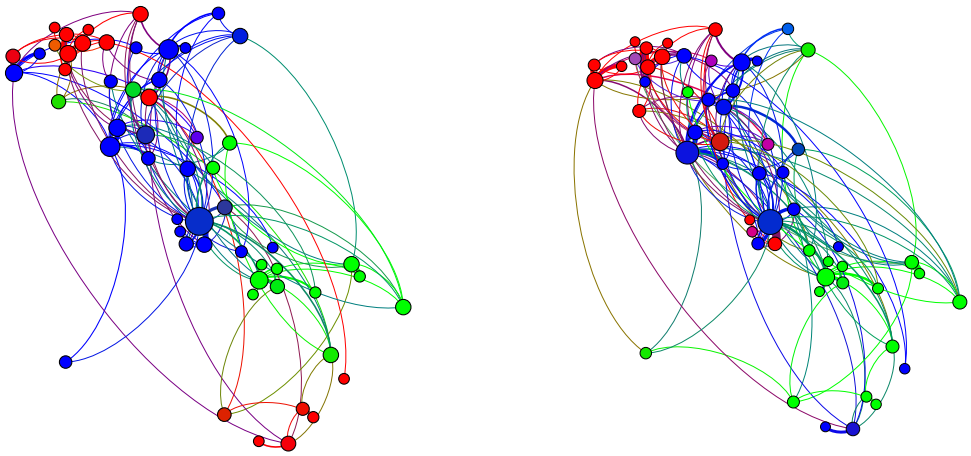
(a) Rete con un livello **(b)** Rete con due livelli

Figura 4.7: Fisica: rete con uno e due livelli tramite la struttura di comunità normalizzata, $K = 2$ gruppi

Nel passaggio alla stima del modello con tre gruppi, si noti da (4.8) che, nuovamente, in un caso ci si allontana da quanto visto per Statistica mentre nell'altro ci si riconduce a una situazione simile, ma con alcune differenze. In figura 4.8a si possono notare tre gruppi così formati: l'area di Milano insieme alle università più vicine e Trento e Bologna si legano alla Sicilia; il Nord-Est è insieme alla Toscana, alle Marche, Roma e Cagliari; infine, il Sud Italia è legato a quattro università abbastanza separate geograficamente tra loro nel Centro-Nord. Anche in questo ambito il gruppo blu unisce alcune delle più importanti università per il grado, anche se la situazione è meno netta rispetto a Statistica perché, come detto nelle analisi descrittive, i nodi sono in generale più piccoli. Si noti che le due università di Torino afferiscono per la prima volta a due diversi gruppi.

Diversa è la configurazione dei tre gruppi nella rete con due livelli (figura 4.8b), la quale si avvicina alla situazione di Statistica. Nel passaggio da uno a due livelli,

il gruppo che ha subito meno modifiche è quello centrale, mentre si è definito un gruppo a nord-ovest e un gruppo che comprende il Sud Italia nel suo complesso.

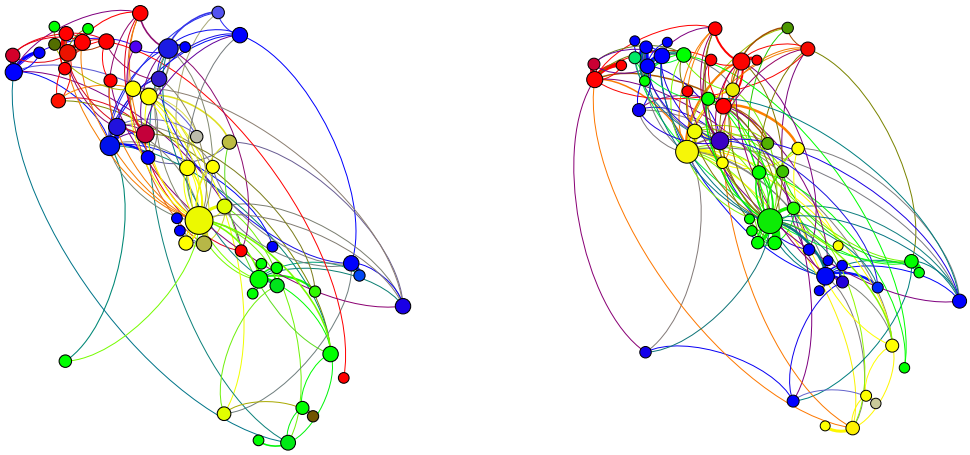


(a) Rete con un livello **(b)** Rete con due livelli

Figura 4.8: Fisica: rete con uno e due livelli tramite la struttura di comunità normalizzata, $K = 3$ gruppi

Da notare che in questa situazione Roma La Sapienza, il nodo più grande, non sembra essere legata alle altre università della città perché alcune fanno parte di un altro gruppo.

Infine, considerando la divisione in quattro gruppi, viene confermato quanto detto in precedenza riguardo la maggiore difficoltà interpretativa del risultato. In questo caso è complesso anche il confronto tra uno e due livelli (figura 4.9). I gruppi stimati sono completamente diversi tra le due configurazioni con pochi aspetti in comune: cambia sia la disposizione geografica generale, sia è completamente diversa la combinazione dei nodi all'interno di ognuno. Un aspetto da sottolineare riguarda il fatto che nella rete con due livelli in generale le università nella stessa città non vengono divise ma appartengono allo stesso gruppo. Questa configurazione non ha collegamenti chiari con l'aspetto geografico perché i vari gruppi non sembrano riunire particolari regioni o aree.



(a) Rete con un livello

(b) Rete con due livelli

Figura 4.9: Fisica: rete con uno e due livelli tramite la struttura di comunità normalizzata, $K = 4$ gruppi

Risulta chiaro come la scelta del numero di gruppi sia fondamentale per l'interpretazione delle reti; in questo caso limite dove i criteri di scelta del K ottimale danno risposte divergenti, visto lo scopo descrittivo della stima del modello, conviene intraprendere un approccio generale e valutare più di una soluzione.

Per concludere, sia nella valutazione delle matrici strutturali, sia dell'interdipendenza tra gli strati, valgono le stesse considerazioni fatte in precedenza perché anche le matrici strutturali stimate per Fisica presentano tutte una struttura assortativa e la stima risulta migliore usando due livelli, come si può vedere in figura 4.10.

4.2 Ranking degli atenei in Italia

Il mercato del lavoro e lo scambio di individui in ambito accademico hanno un ruolo fondamentale nel modellare molti aspetti della vita accademica, quali la qualità della ricerca nel suo complesso e i risultati educativi. L'ipotesi implicita che

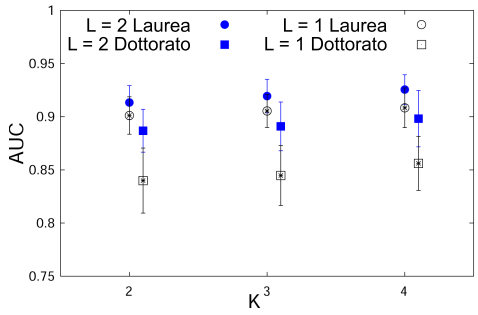


Figura 4.10: Valori dell'AUC per $K = 2, 3, 4$ per valutare l'interdipendenza degli strati in Fisica

si fa quando si considera l'assunzione presso un ateneo di un individuo proveniente da uno diverso è che l'ateneo che assume sta valutando positivamente la formazione e i programmi di ricerca della prima università. D'altro canto, quando un individuo accetta un lavoro presso una certa istituzione, dà una valutazione positiva sulla qualità di questa (Clauset, Arbesman e Larremore 2015).

È naturale aspettarsi una differenza del tasso di successo nel posizionamento dei propri laureati o dottori di ricerca tra le università, dovuta alla dominanza di alcuni atenei su altri o, detto in altro parole, alla presenza di una gerarchia sociale generata dal prestigio. Bisogna sottolineare che il prestigio è costituito sia da aspetti di differenza nel merito accademico, sia da una parte non meritocratica come lo status sociale o per esempio il posizionamento geografico. L'interesse, quindi, è nello stilare un ordinamento delle università, utilizzabile per esempio dallo Stato per erogare fondi in modo proporzionale al prestigio o anche da futuri fruitori di un determinato corso di studi per valutarne il posizionamento.

Si sottolinea, prima di iniziare con le analisi, che verrà trattato per semplicità solo il caso con un solo livello costruito dall'istruzione maggiore.

Come visto nelle analisi descrittive, la formazione di individui nelle università è molto asimmetrica in quanto il 44% degli atenei per statistica e il 30% per fisica non ha nessun docente in uscita. In più, anche se è massiccia la presenza di auto-archi, essi non vi sono in tutte le università e solo il 45% degli atenei di statistica e il 64% di fisica li presenta. Infine, per quanto detto sulla forza in entrata del nodo, che si può vedere come approssimazione della grandezza del dipartimento all'interno dell'ateneo, si ha una terza asimmetria sulla grandezza degli atenei

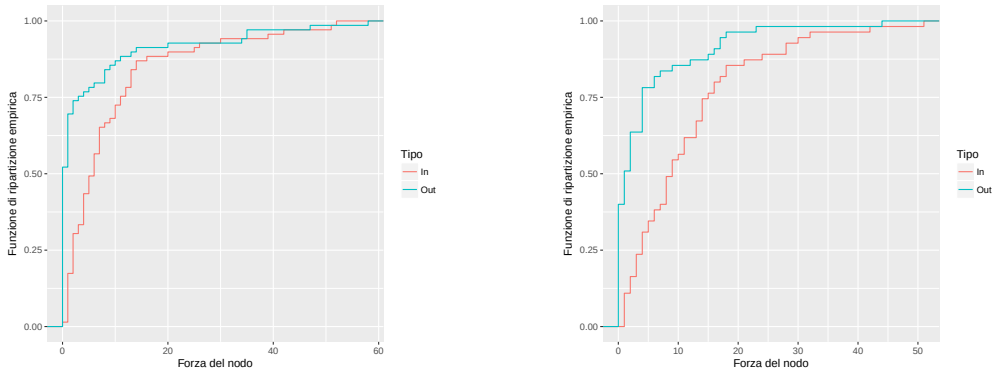
dovuta al range molto ampio di valori che essa assume nei dati (si veda tabella 4.3 per alcune misure riassuntive).

	Statistica	Fisica
Nodi	69	55
Archi	629	630
Auto-archi assenti	55%	36%
Media forza out	5.0	4.4
Range forza out	0-58	0-44
Forza out assente	44%	30%
Media forza in	9.1	11.5
Range forza in	0-52	0-51

Tabella 4.3: Alcune misure riguardo la forza dei nodi per settore

Si potrebbe pensare che il posizionamento dei docenti sia proporzionale alla grandezza dell'ateneo e questo, se fosse vero, porterebbe ad avere una gerarchia solo di tipo dimensionale per ogni settore disciplinare e quindi non avrebbe senso un algoritmo per estrarre un ranking. Risulta quindi fondamentale valutare tramite un test di Kolmogorov-Smirnov (KS) se capacità di posizionamento e grandezza dell'ateneo si distribuiscono in modo statisticamente differente, che in altre parole significa applicare un test KS tra *out-strength* e *in-strength*. Si noti che nel secondo caso vengono contati gli auto-archi perché se si parla di produzione bisogna tenere conto anche degli individui formati che rimangono interni all'ateneo (questa distinzione vale anche per le misure presentate in tabella 4.3). Anche se le forze in entrata e uscita sono correlate positivamente tra loro in entrambi i casi (entrambe le correlazioni sono circa 0.7), i test KS per Statistica e Fisica rifiutano l'ipotesi nulla di uguaglianza delle due distribuzioni, indicando un differenziale nel posizionamento dei docenti che non deriva meramente dalla grandezza delle stesse (in figura 4.11 si veda il confronto tra le ripartizioni empiriche).

Dall'altro lato, è interessante valutare una misura della disuguaglianza sociale, che si può calcolare tramite l'indice di Gini sulla produzione dei docenti per i due settori, quindi sulla forza in uscita. Entrambi i settori presentano una situazione di disuguaglianza e in particolare $G_{stat} = 0.79$ e $G_{fis} = 0.65$, con $G = 1$ che denota disuguaglianza perfetta. In figura 4.12 si possono confrontare le curve disegnate



(a) Statistica **(b)** Fisica

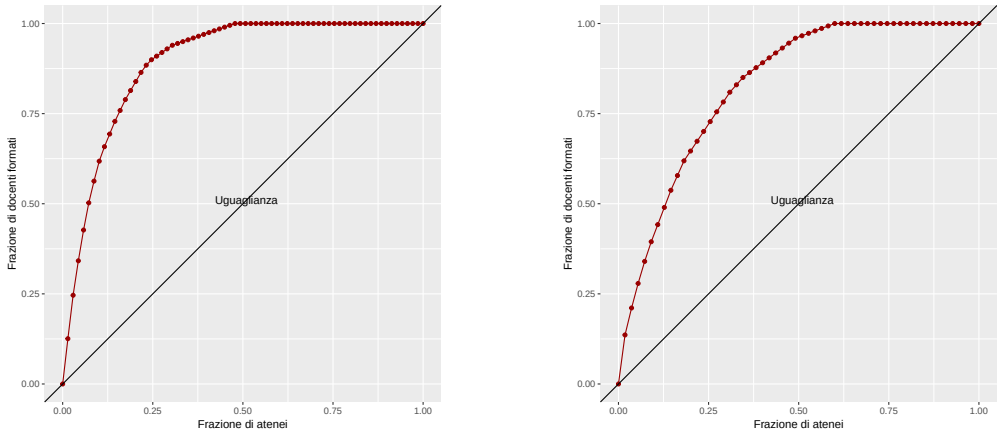
Figura 4.11: Funzione di ripartizione empirica di forza in entrata e uscita per entrambi i settori disciplinari

dagli atenei con la diagonale, che rappresenta la situazione di uguaglianza tra frazione di atenei e frazione di docenti formati, confermando la non presenza di equità. Solo il 25% di atenei in Statistica producono quasi il 90% dei futuri docenti; in Fisica la percentuale di individui formati si abbassa al 75% ma rimane comunque alta rispetto ai valori della diagonale.

Prima di procedere con il ranking, per assicurarsi che non ci siano limitazioni nel calcolo, si misura la reciprocità tra i nodi e cioè si valuta in che proporzione sono presenti nodi che si scambiano lo stesso numero di archi. Formalmente, si valutano tutte le coppie di celle della matrice di adiacenza A_{ij} e A_{ji} e si calcola il rapporto

$$R = \frac{\min A_{ij}, A_{ji}}{\max A_{ij}, A_{ji}} \quad \text{con } i, j \in \{1, ..., N\}$$

che è sempre un numero positivo tra zero e uno, dove lo zero denota uno squilibrio e quindi la presenza di uno dei due nodi dal quale non parte nessun arco e uno indica una situazione di reciprocità dove i e j sono connessi dallo stesso numero di archi. Naturalmente, si tralascia il calcolo per le coppie di nodi con entrambe le celle nulle della matrice di adiacenza e per la diagonale della matrice che avrà sempre valore uno. Si crea in questo modo una nuova matrice simmetrica che in entrambi i settori di studio ha un tasso di nodi reciproci pari circa all'1% del totale. Il calcolo di questo valore è importante perché tra due atenei che si scambiano lo stesso



(a) Statistica **(b)** Fisica

Figura 4.12: Curva di Lorenz della forza in uscita rispetto alla frazione di atenei per entrambi i settori disciplinari

numero di individui, l'algoritmo che calcola il ranking non saprebbe posizionare le due università e il ranking risulterebbe più incerto.

Si presentano nel seguito i risultati dell'applicazione dell'algoritmo su entrambi i dataset, precisando che il ranking è possibile per un numero ridotto di atenei. Infatti, per i nodi che hanno grado in entrata o in uscita nullo senza contare gli auto-archi non è possibile il calcolo di un ordinamento, perché non hanno flussi in entrambe le direzioni ed è invece l'uso congiunto di queste che permette di valutare il posizionamento dei nodi in graduatoria. Inoltre, si considera anche il caso dell'algoritmo pesato perché si ritiene più corretto valutare in modo sempre maggiore le violazioni di più posizioni in graduatoria.

4.2.1 Statistica

Grazie all'algoritmo applicato ai dati a disposizione, si possono ottenere diverse informazioni sul ranking calcolato. In particolare, oltre all'ordinamento medio o mediano sul quale si possono estrarre le considerazioni finali, viene stimata anche la distribuzione ottenuta tramite diecimila iterazioni dell'algoritmo. In figura 4.13 si possono visualizzare nel loro complesso queste informazioni per il settore statistica, nel caso non pesato.

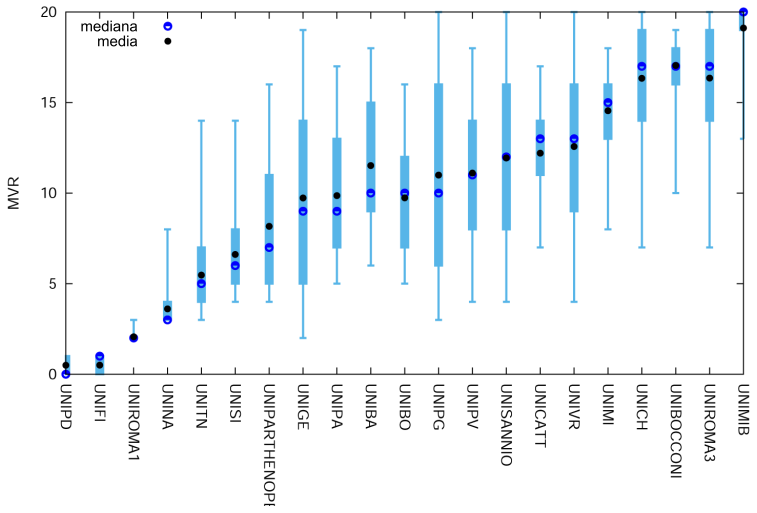


Figura 4.13: Classifica degli atenei di Statistica tramite Minimum Violation Ranking

Innanzitutto si noti che il numero di atenei ordinati è pari a 21, i quali sono gli unici atenei che non presentano grado nullo contemporaneamente in entrata e in uscita. Il grafico va letto dall'angolo in basso a sinistra, dove si trova la prima posizione del ranking, all'angolo in alto a destra. Bisogna sottolineare che non vengono eliminate solo le università di cui si è parlato nelle analisi descrittive, ma, dopo una prima scrematura, viene ricalcolato il grado di ogni ateneo rimasto ed eliminati nuovamente quelli con grado zero. Dopo due controlli di questo tipo, rimangono università con grado in entrata e in uscita positivi, dati dai collegamenti con gli altri atenei rimasti.

Le statistiche fondamentali delle distribuzioni empiriche di ognuno degli atenei ordinati sono rappresentate tramite boxplot. Grazie a essi, si può cogliere la differenza molto marcata tra alcuni atenei che presentano una distribuzione molto schiacciata, che denota la poca variabilità della posizione di quell'ateneo nelle replicazioni dell'algoritmo, e altri con una posizione molto più variabile.

Nelle prime posizioni si trovano le università di Padova, Firenze e Roma La Sapienza con ognuna una posizione poco variabile nelle diverse replicazioni.

In generale tutte le università di cui in precedenza si è parlato nella descrizione dei gradi sono presenti nel ranking, anche se con posizioni spesso completamente diverse. Riguardo le prime posizioni nel ranking, si può dire che siano in linea con quanto ci si poteva aspettare dalle precedenti analisi. L'università di Firenze, che si

trova terza nel grado totale e in uscita e riguardo la forza è presente come quinta in quella totale, nell'ordinamento risulta seconda. Padova e Roma La Sapienza sono ben posizionate in tutte le classifiche stilate sulle tipologie dei gradi e anche nell'ordinamento mantengono due posizioni tra i primi atenei.

Riguardo l'appartenenza geografica delle università nel ranking, sono rappresentate tutte le tre macro-regioni geografiche. Sono presenti 10 atenei del Nord, 4 del centro Italia e 7 del Sud, denotando una proporzione maggiore della zona nord rispetto a quelle generali che si hanno in Statistica.

Tra gli atenei mega, nel ranking mancano Torino, Pisa e Catania in quanto risultano con grado nullo dopo aver tolto il primo gruppo di università. Nonostante la presenza di quasi tutte le università di questa dimensione, non si trovano tutte tra le prime posizioni, a sottolineare di nuovo che la grandezza come numero di iscritti non assicura in generale una maggior importanza nelle analisi svolte in questo lavoro.

Come detto, considerare violazioni di pari peso anche se tra due posizioni molto lontane della classifica, sembra meno intuitivo rispetto al pesarle per la differenza di posizioni. Per questo motivo si è stimato anche un ordinamento degli atenei con l'algoritmo pesato (si veda figura 4.14).

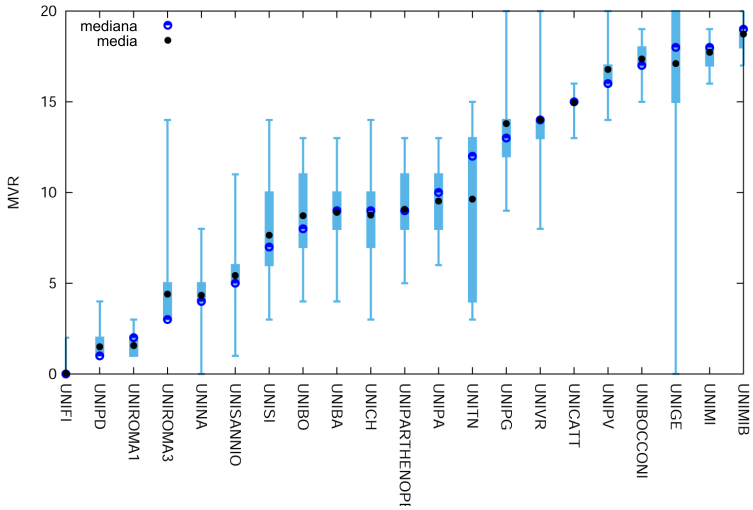


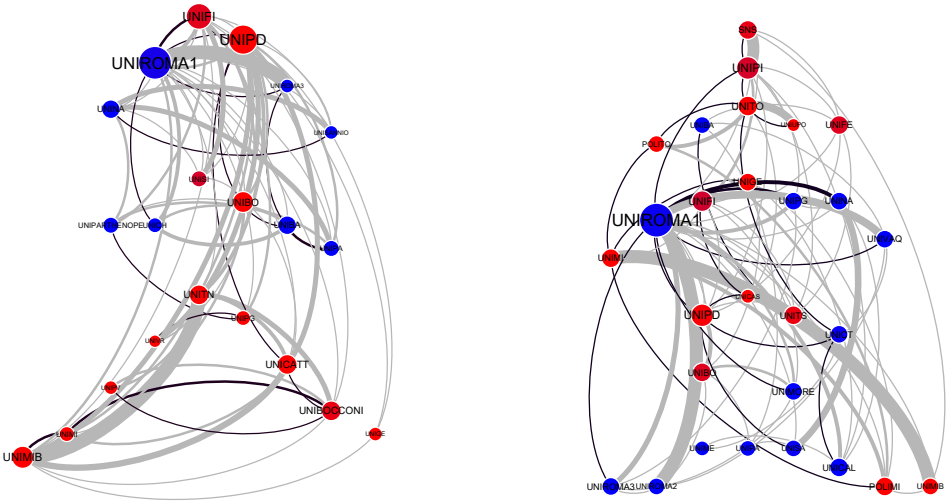
Figura 4.14: Classifica degli atenei di Statistica tramite Minimum Violation Ranking pesato

Si nota, innanzitutto, una generale minor variabilità nelle distribuzioni empi-

riche; rimangono però diversi valori estremi che allungano molto alcune code dei boxplot.

In secondo luogo, c'è un cambiamento anche nelle prime posizioni della classifica: l'università di Firenze si posiziona prima nel caso pesato. Si hanno molte differenze tra i due ordinamenti e ci sono università che cambiano completamente posizione usando un diverso algoritmo: per esempio, Roma Tre passa dalla penultima posizione alla quarta posizione nel caso pesato o invece, Trento, citata in precedenza nelle analisi descrittive, perde posizioni se si considera la versione pesata.

È possibile visualizzare questi ranking anche sotto forma di rete, che permette di vedere anche le violazioni verso l'alto. Tramite l'output dell'algoritmo e la visualizzazione tramite Gephi, si può creare una rappresentazione quale (4.15), dove si possono unire diversi aspetti visti sia in questa sezione in particolare sull'algoritmo pesato, sia nella precedente riguardante la *community detection*.



(a) Statistica **(b)** Fisica

Figura 4.15: Visualizzazione tramite ranking pesato delle reti con colori relativi alla stima del modello, $K = 2$ gruppi

Soffermandosi inizialmente in figura 4.15a, si visualizzano le posizioni degli

atenei in graduatoria a partire dall'alto della figura. Bisogna prestare attenzione in particolare a quei nodi posizionati circa alla stessa posizione in senso orizzontale perché sono gli atenei che hanno una posizione simile nell'ordinamento secondo l'algoritmo pesato.

Riguardo i nodi, la dimensione è ancora una volta proporzionale al grado, una volta sottratti i collegamenti con le università eliminate dal ranking. Il colore, invece, riprende la suddivisione stimata dal modello con due gruppi e un livello della rete. Grazie a esso, si può vedere che la maggior parte dei nodi rimasti sono di colore rosso, che rappresenta come detto il gruppo del Nord (per il confronto con il risultato del modello si veda figura 4.3a). Naturalmente, questo confronto può essere riproposto con le diverse numerosità di gruppo stimate dal modello e aiuta l'interpretazione congiunta dei due risultati presentati nella tesi.

Riguardo i collegamenti tra i nodi, sono stati colorati tramite il colore nero gli archi che presentano una violazione, cioè quelli direzionati verso una posizione più alta del ranking rispetto a quella del nodo da cui partono, mentre tutti gli altri sono grigi e puntano quindi correttamente verso il basso della graduatoria. Anche in questo caso hanno uno spessore proporzionale al peso dell'arco nella rete, eliminando prima gli archi collegati a nodi non utilizzati nella stima del ranking.

Una volta descritto l'ordinamento ottenuto, è importante poterlo confrontare con uno ufficiale, che permetta di affermare se quanto ottenuto rispecchi il prestigio nel senso di qualità e merito accademico di un ateneo.

In Italia, l'unica classifica ufficiale utilizzata dallo Stato e dal MIUR per la ripartizione dei fondi agli atenei è stilata da ANVUR (Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca, VQR 2011-2014) e viene fatta tramite una valutazione della ricerca che si ottiene tramite indicatori che giudicano prodotti di ricerca (articoli, monografie, atti di convegni, software, etc.) in base a "criteri di rilevanza, originalità e grado d'internazionalizzazione con due diverse metodologie, valutazione bibliometrica e peer review". Riassumendo, si tratta di un processo valutativo basato su un criterio di merito. Proprio per questo motivo, si opera un confronto tra il ranking ottenuto dall'algoritmo e quanto pubblicato da ANVUR.

Vanno fatte alcune precisazioni riguardo a questo confronto. Come detto, l'ordinamento ottenuto tramite l'algoritmo per Statistica riguarda solo 21 atenei per problemi strutturali dei dati disponibili; al contrario, la classifica di ANVUR è

formata da tutti gli atenei che presentano almeno un prodotto della ricerca nel determinato macro-settore o settore disciplinare che si considera. Si noti, innanzitutto, che questa può essere una criticità delle classifiche di ANVUR perché vengono stilate indipendentemente dal numero di lavori inviati dalle università, perciò possono esserci anche università il cui numero di lavori valutati è molto esiguo.

Per evitare il confronto tra vettori contenenti atenei completamente diversi si sono estratti dalle classifiche stilate da ANVUR solamente gli atenei selezionati dall'algoritmo, con i corrispondenti punteggi associati. In questo modo si confrontano vettori formati dagli stessi atenei, posizionati in modo diverso e si cerca di valutare l'estensione di questa diversità. Per Statistica si sono utilizzate le classifiche generale dell'area 13, Scienze Economiche e Statistiche, e per l'insieme di settori disciplinari utilizzati in questo lavoro.

La correlazione, si veda tabella 4.4, è stata calcolata tramite la misura di *Spearman* per gli ordinamenti veri e propri degli atenei e tramite R^2 *di Pearson* per i vettori dei punteggi associati agli atenei (nel caso delle classifiche ANVUR viene usata una misura della qualità detta "voto medio normalizzato" mentre nel caso dell'algoritmo si usa la media del MVR).

Si nota una correlazione negativa, anche se piuttosto bassa e non significativa, tra il MVR ed entrambe le classifiche ANVUR, mentre, come ci si poteva aspettare, le correlazioni tra le due classifiche ANVUR sono positive, molto alte e decisamente significative. Si noti che la non significatività porta a non poter trarre delle conclusioni in modo deciso, ma a delineare solamente una linea generale.

La correlazione di *Spearman* misura la differenza tra i vettori sulla base dei cambiamenti di posizione che bisogna operare per renderli uguali e conferma che gli atenei hanno posizioni molto diverse tra le classifiche. L'R^2 *di Pearson* misura

	MVR	Gen (A)	Sett (A)	MVR	Gen (A)	Sett (A)
MVR	1.00			1.00		
Gen (A)	-0.37(0.103)	1.00		-0.33(0.138)	1.00	
Sett (A)	-0.40(0.073)	0.82($< 10^{-5}$)	1.00	-0.40(0.071)	0.85($< 10^{-5}$)	1.00

Tabella 4.4: Matrici di correlazione per Statistica tra MVR e due classifiche stilate da ANVUR (generale e ristretta ai settori disciplinari): a sinistra *Spearman*, a destra *Pearson* (tra parentesi i *p-value*)

la correlazione lineare tra i vettori dei punteggi; in entrambe le misure il segno negativo sottolinea l'andamento opposto dei vettori dei ranking.

Ciò che si può affermare dopo la valutazione dei risultati, tenendo conto che risultano non significativamente diversi da zero, è che la classifica che si ottiene tramite lo studio della rete, che perciò rispecchia il sistema di valutazioni implicite operate da docenti e atenei, non rispecchia invece una gerarchia puramente di merito ma anzi sono predominanti gli altri fattori, che allontanano il ranking calcolato da quello ufficiale basato sul merito.

Riguardo a questo risultato così contrastante rispetto al caso americano dove ranking ufficiale ed estratto dall'algoritmo risultavano molto correlati, si noti che il numero di osservazioni utilizzate nel caso italiano risulta molto basso e questo è aggravato anche dall'ulteriore selezione che si è dovuta operare prima dell'uso dell'algoritmo.

Uno spunto interessante per la futura ricerca potrebbe essere in un confronto ristretto, tra il ranking ottenuto e ANVUR, solo di atenei i quali congiuntamente abbiano un numero sufficiente di archi nella rete e di prodotti valutati e possibilmente che non facciano parte di categorie particolari di atenei come i più recenti. Prime evidenze esplorative sembrano mostrare che tra un insieme di università più solido sotto questi punti di vista, possa esserci un'evidenza meno marcata di distanza tra il ranking ottenibile e ANVUR.

4.2.2 Fisica

Lo stesso procedimento appena descritto viene ripetuto per il settore Fisica.

Innanzitutto, valutando la classifica ottenuta tramite l'algoritmo non pesato (figura 4.16), si nota subito una maggior variabilità delle stime e quindi una maggior incertezza della classifica stilata rispetto a Statistica. Questo è sicuramente dovuto alla minor quantità di dati nel dataset di Fisica perché, si ricorda, solo il 40% degli archi consiste in spostamenti tra università diverse mentre il restante riguarda docenti che rimangono nella stessa università e quindi non partecipano al calcolo del ranking.

Altra differenza rispetto al caso precedente riguarda il numero di atenei rimasto nel calcolo che sale in questo caso alla metà dei nodi, 28; era infatti minore il

numero di università senza archi in uscita e nullo quello di università senza archi in entrata.

Al primo posto del ranking troviamo la Scuola Normale Superiore; altre università citate in precedenza nelle classifiche per grado o forza come Roma La Sapienza, Milano e Pisa si trovano nelle prime posizioni mentre in altri casi l'ordine cambia, ad esempio Padova si sposta in questo caso a circa metà classifica.

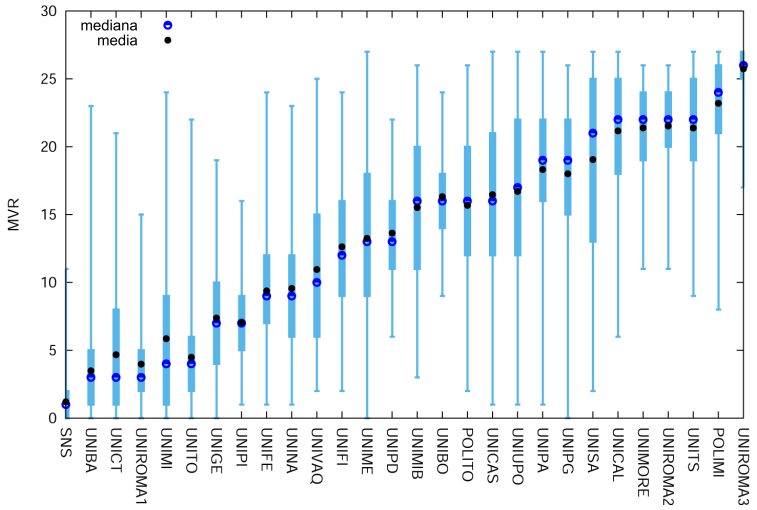


Figura 4.16: Classifica degli atenei di Fisica tramite Minimum Violation Ranking

Incrociando l'informazione ottenuta tramite il ranking con le informazioni aggiuntive raccolte sugli atenei, si può valutare che poco meno della metà (il 43%) degli atenei presenti nell'ordinamento per Fisica fa parte del macro-gruppo del Nord Italia, e circa il 29% del Centro e del Sud. Vengono quindi riprese circa le stesse proporzioni dello studio generale per Fisica per Nord e Centro, ma il Sud è sotto-rappresentato. Inoltre, in questo ranking sono presenti tutte le università mega citate in precedenza.

Se si passa alla valutazione della figura 4.17, che rappresenta il ranking pesato, si può dire che la variabilità è lievemente diminuita e, come nel caso di Statistica, la maggior parte delle università ha cambiato posizione.

Al contrario di quanto riscontrato nel caso di Statistica, nel passaggio tra le due modalità di calcolo del ranking, cambiano quasi totalmente le prime tre posizioni. A parte la Scuola Normale Superiore, che rimane prima, la seconda e terza posizione

vengono assegnate in un caso a Bari e Catania, nell'altro a Pisa e Torino, spostando Bari nella posizione successiva e Catania dopo la metà della classifica.

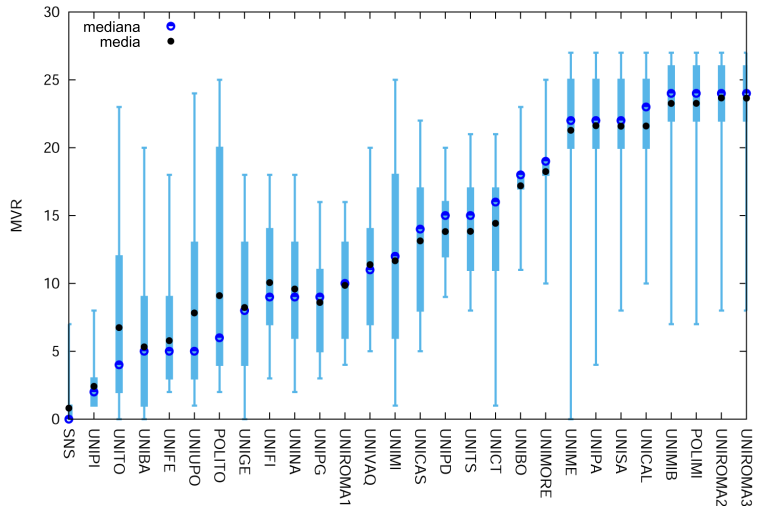


Figura 4.17: Classifica degli atenei di Fisica tramite Minimum Violation Ranking pesato

Considerando a questo punto figura 4.15b, è utile valutare inizialmente il settore Fisica nello specifico, successivamente anche un confronto tra i due ambiti disciplinari.

In questa rappresentazione sono stati riportati i colori per la suddivisione in due gruppi stimata dal modello, anche se come detto non era condivisa da entrambi i criteri di scelta. Si può riprodurre lo stesso grafico con le altre numerosità di gruppo, sulla base dell'interpretazione finale che si sceglie di dare e affiancando quindi a una rappresentazione geografica usata per valutare gli spostamenti dei docenti, una figura riportante il ranking per capire come si posizionano gli atenei.

Si nota un bilanciamento tra i nodi blu e i nodi rossi e la presenza di numerosi nodi affiancati orizzontalmente.

Nel confronto, si nota che il nodo di Roma La sapienza è il più grande come in Statistica ma, mentre nel caso precedente sono presenti altri nodi di circa la stessa dimensione o grado, in Fisica tutti gli altri atenei sono più piccoli. Riguardo a questo, è interessante vedere che Roma La Sapienza, nonostante la grande dimensione, non si trova nelle prime posizioni del ranking.

Le prime tre posizioni sono ordinate in modo più netto rispetto al caso precedente e si può dire che la proporzione di violazioni sul totale degli archi è circa la stessa in entrambi i settori.

Allo stesso modo, si può operare il confronto con il ranking ufficiale anche per il settore Fisica. In questo caso si sono utilizzate la classifica generale per l'area 02, Scienze Fisiche, e quella specifica del settore disciplinare FIS01. Un problema riscontrato solo in questo settore è la mancanza di uno degli atenei classificati tramite MVR nelle classifiche ufficiali e si tratta dell'Università degli Studi di Cassino. Per questo motivo le misure di confronto sono state ottenute con 27 atenei.

In tabella 4.5 si nota che la correlazione fra le due tipologie di ranking cambia di segno rispetto al caso precedente, a denotare una maggior corrispondenza in questo settore rispetto a Statistica, e la forza della relazione è maggiore quando ci si restringe alla fisica sperimentale portando a conclusioni significative.

Anche se rimane valida l'interpretazione spiegata nel paragrafo precedente sul fatto che il ranking ottenuto non rispecchia il merito, l'evidenza è meno forte. Si noti che questo aspetto va ad aggiungersi alle differenze evidenziate in precedenza tra Statistica e Fisica e conferma la scelta iniziale di confrontare queste due materie, assumendo sia alcuni aspetti di somiglianza ma anche altri di contrasto.

	MVR	Gen (A)	Sett (A)	MVR	Gen (A)	Sett (A)
MVR	1.00			1.00		
Gen (A)	0.35(0.017)	1.00		0.37(0.014)	1.00	
Sett (A)	0.52(0.008)	0.80($< 10^{-5}$)	1.00	0.50(0.005)	0.82($< 10^{-5}$)	1.00

Tabella 4.5: Matrici di correlazione per Fisica tra MVR e due classifiche stilate da ANVUR (generale e ristretta al settore disciplinare): a sinistra *Spearman*, a destra *Pearson* (tra parentesi i *p-value*)

Capitolo 5

Conclusioni

Questo lavoro segue l'approccio proposto da Clauset, Arbesman e Larremore (2015) nello studio della mobilità dei docenti universitari in diversi settori disciplinari. Nello specifico, si ripropongono per l'Italia le analisi svolte originariamente in tre settori (*Computer science*, *Business* e *History*) su 19 000 docenti che lavorano in università del Nord America. I dati usati in questa tesi riguardano due settori disciplinari: Statistica e Fisica sperimentale, i quali hanno in comune la forte propensione per la ricerca, ma mostrano anche alcuni aspetti di contrasto.

Nel complesso, i dati utilizzati si compongono di rispettivamente 69 e 55 atenei italiani per Statistica e Fisica sperimentale, i quali contengono 629 e 630 docenti con le informazioni minime per non essere considerati dati mancanti. Entrambi gli insiemi di dati sono stati raccolti manualmente, dopo l'estrazione della lista dei docenti dall'archivio CINECA, tramite la ricerca dei *curricula* dei docenti dai siti istituzionali delle università o da altre fonti.

Sono stati eliminati i docenti per i quali l'informazione sulla formazione (laurea e dottorato di ricerca) era completamente mancante e anche coloro che hanno ottenuto entrambi i titoli all'estero. L'informazione riguardo l'estero risulta quindi mancante sia per la scelta di eliminare i docenti formati all'estero, sia per la porzione di docenti italiani che hanno scelto di lavorare fuori dall'Italia, data la mancanza di questa informazione nella lista iniziale.

Si è visto come sia possibile estrarre numerose informazioni di interesse esaminando i dati da un punto di vista statico, tramite lo studio di docenti e atenei separatamente. Per entrambe le unità di interesse si sono usate diverse informazio-

ni atte a descriverle, quali per esempio il genere e il ruolo ricoperto per i docenti o la macro-regione geografica e la tipologia per gli atenei.

È chiaro come il considerare i dati anche da un punto di vista dinamico possa fornire diversi livelli interpretativi e arricchire l'analisi. Per questo motivo sono state costruite due matrici di adiacenza per ognuno dei settori, tramite le quali sono state stimate due reti: la prima considera l'istruzione maggiore e quindi prende, se disponibile, l'informazione sul dottorato, altrimenti quella sulla laurea, mentre la seconda utilizza la specificazione in due strati separati dei due titoli di studio.

Sono stati portati avanti due obiettivi in modo parallelo: applicare una *community detection* ai dati per individuare delle comunità latenti e creare un ordinamento tra gli atenei che fosse generato dalla struttura della rete. Per quanto riguarda il primo obiettivo è stato utilizzato lo *stochastic blockmodel* multilivello per comunità miste, mentre per il secondo si è considerato un algoritmo basato sul *minimum violation ranking*.

I risultati ottenuti, pur evidenziando diverse criticità con riferimento alla specifica applicazione considerata, hanno comunque permesso di fare alcune considerazioni di interesse.

In particolare, riguardo il primo obiettivo, si è potuta ottenere una descrizione dei gruppi stimati dal modello, anche con diverse numerosità in modo da ottenere un confronto. Si è ritenuto importante visualizzare la rete tramite il posizionamento geografico degli atenei in modo da valutare, di caso in caso, se fosse possibile un'intuitiva interpretazione geografica della suddivisione in gruppi. Il confronto tra le due versioni di reti costruite tramite le matrici di adiacenza ha permesso di valutare informazioni diverse riguardo il titolo di studio.

Le conclusioni sono di una dipendenza geografica abbastanza netta per Statistica, dove l'utilizzo dei due titoli di studio separati porta a confermare le intuizioni che si ottengono tramite l'istruzione maggiore, sia nel caso di due che di tre gruppi. Interessante è notare che il terzo gruppo nel caso dei due livelli ha una connotazione diversa rispetto a quella geografica, perché infatti coglie aspetti di importanza per la rete di tre atenei distribuiti dal Nord al Centro Italia.

Per quanto riguarda il settore Fisica, non si riescono a trarre conclusioni che vadano verso una chiara motivazione puramente geografica degli spostamenti, soprattutto per le differenze marcate tra le specificazioni con uno e due livelli. In particolare, si giunge alle stesse conclusioni di Statistica se si considerano i due

strati, mentre il livello di istruzione maggiore fornisce un'informazione diversa. Inoltre, la scelta del numero di gruppi per questo ambito non è così chiara come il caso precedente e anche questo denota una situazione meno netta; si noti come questo potrebbe rispecchiare anche il minor numero di dati a disposizione riguardo agli spostamenti tra atenei.

Per ciò che concerne il ranking tra gli atenei, è risultato chiaro dopo alcune analisi l'utilità nell'estrarre un ordinamento che rispecchiasse la gerarchia intrinseca della struttura di rete nei dati. Infatti, essa risulta slegata da concetti quali la grandezza degli atenei o la capacità di formazione. Una classifica del genere tra università può trovare una sua utilità da un lato per la possibilità per lo Stato di erogare fondi in modo proporzionale all'ordinamento e dall'altro per la possibilità per i futuri fruitori di un determinato corso di studi di valutare l'ateneo migliore. Inoltre, si possono unire i due aspetti visti di *community detection* e ranking per una lettura globale dei risultati.

La versione pesata del ranking sembra fornire risultati migliori vista la minor variabilità delle stime nelle replicazioni dell'algoritmo. In entrambi i settori disciplinari le prime posizioni del ranking pesato rispecchiano alcune idee che si potevano trarre dalla prima analisi esplorativa dei dati; in generale sono presenti tutte le università di cui si era parlato nei diversi punti delle analisi descrittive, ma bisogna sottolineare che l'ordinamento si è potuto ottenere solamente per un sottoinsieme di atenei, i quali presentavano contemporaneamente grado in entrata e in uscita positivi.

Per valutare se effettivamente il ranking estratto dalla rete rappresentasse una classifica del merito accademico o qualità delle università, si è operato un confronto di questo con la classifica stilata da ANVUR, basata sulla qualità della ricerca e usata dallo Stato per ripartire i fondi. Mentre l'assunzione sul prestigio derivante dal merito viene rispettata dalle classifiche ufficiali, in quanto sono costruite proprio come valutazione dei prodotti della ricerca delle università, si è visto come il ranking estratto dalla rete non la rispetti. Si pensa che questo da un lato possa dipendere dalle basse numerosità dei dati a disposizione, che non permettono uno studio corposo dei flussi tra università, mentre dall'altro rispecchi una caratteristica propria degli spostamenti dei docenti italiani tra gli atenei. Essi infatti, tolta l'importante fetta di quelli che cercano di rimanere a lavorare nella stessa università della propria formazione, si spostano non per motivi di prestigio e merito ma

apparentemente a causa di altri fattori quali per esempio la posizione geografica. Il ranking così costruito rispecchia sì la rete, ma non si può dire che con questa rete si possa inferire sul prestigio inteso nel senso di qualità degli atenei.

D'altro canto, sono diverse le criticità riscontrate, che derivano soprattutto dalla scelta e dalla costruzione dei dati e influenzano entrambi i temi affrontati in questa tesi.

Innanzitutto è chiara la differenza nella quantità di dati nel confronto tra la situazione americana e quella italiana. Le numerosità sono nettamente inferiori anche considerando interi macro-settori, ma in questo lavoro è stata fatta un'ulteriore riduzione a singoli settori disciplinari e ciò porta a un minor numero di docenti considerati e quindi a problemi legati a questo aspetto.

Oltre a questa numerosità ridotta iniziale, si sono rilevati un tasso non esiguo di valori mancanti, dovuti alla difficoltà di reperimento di informazioni di taluni docenti.

Altra caratteristica propria dell'Italia è la presenza del dottorato solo dal 1980 e ciò spiega le scelte attuate riguardo alle due costruzioni delle matrici di adiacenza, ma è ovvia la differenza dal caso americano dove si può valutare la completa informazione sul Ph.D., tralasciando quella sulla laurea. Si pensa, infatti, che l'informazione sul dottorato possa essere migliore per una ricerca in questo ambito perché in generale la scelta stessa di conseguire questo titolo è legata alla propensione della persona alla ricerca e si può assumere che ci sia dietro un ragionamento sull'ateneo in merito.

Come detto in precedenza, un ulteriore problema dei dati riguarda l'estero e le scelte che si sono dovute attuare in merito a questa mancanza della lista iniziale. Mentre per la costruzione del dataset americano si è iniziato da un insieme di facoltà per poi estrarre da ognuna la lista dei docenti, la procedura nel caso italiano ha seguito il senso opposto, essendo iniziata con la ricerca dei docenti di determinati settori disciplinari.

Un aspetto critico non derivante dai dati ma piuttosto da una particolarità culturale italiana, è la predominanza dei docenti che scelgono di rimanere, dopo la laurea o il dottorato, nello stesso ateneo o che scelgono di tornarci. Ciò inficia lo studio della mobilità perché obbliga a considerare solamente un sottoinsieme dei dati a disposizione.

Per rispondere a queste criticità, diverse sono le strade percorribili e gli spunti

per la futura ricerca. Per quanto riguarda i dati, si potrebbero riproporre le stesse analisi a interi macro-settori per avere una numerosità maggiore e valutare se e come cambiano i risultati. Inoltre, una ricerca più approfondita riguardo i dati mancanti, potrebbe abbassarne ulteriormente il tasso e migliorare quanto fatto. Infine, un aspetto non valutato in questa tesi riguarda la possibilità di seguire la carriera di ogni docente dal contratto di ricercatore in poi, valutando anche le tipologie di contratto precedenti, gli anni tra il titolo ottenuto e il primo lavoro e aspetti analoghi.

Dal punto di vista metodologico, invece, è di estremo interesse adeguare modelli e algoritmi a un contesto come quello italiano dove la permanenza nello stesso ateneo rappresenta la norma invece di un'eccezione, quale era nel caso americano. Soprattutto l'algoritmo di ranking dovrebbe poter tenere conto in modo più corposo degli auto-archi per meglio leggere la struttura del sistema universitario italiano.

Bibliografia

Airoldi, E.M. et al. (2008). «Mixed membership stochastic blockmodels». In: *Journal of Machine Learning Research* 9, pp. 1981–2014.

Ball, B., B. Karrer e M.E.J. Newman (2011). «Efficient and principled method for detecting communities in networks». In: *Physical Review E* 84.3. URL: https://doi.org/10.1103/PhysRevE.84.036103.

Bastian, M., S. Heymann e M. Jacomy (2009). *Gephi: An Open Source Software for Exploring and Manipulating Networks*. URL: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

BNCF. *Biblioteca Nazionale Centrale di Firenze, Catalogo Online*. Ultima visita il 09/02/2017. URL: http://www.bncf.firenze.sbn.it/.

Bush, A. (2016). «Mobilità dei docenti universitari in Italia: un'analisi sui flussi dall'Ateneo di laurea a quello d'insegnamento». Tesi di laurea magistrale. Università degli Studi di Padova.

Clauset, A., S. Arbesman e D.B. Larremore (2015). «Systematic inequality and hierarchy in faculty hiring networks». In: *Science Advances* 1.1. URL: 10.1126/sciadv.1400005.

CINECA. *Consorzio Interuniversitario senza scopo di lucro, Cerca Università*. Ultima visita il 08/02/2017. URL: http://cercauniversita.cineca.it/.

De Bacco, C. et al. (2017). «Community detection, link prediction, and layer interdependence in multilayer networks». In: *arXiv preprint*. URL: arXiv:1701.01369.

De Vries, H. (1998). «Finding a dominance order most consistent with a linear hierarchy: a new procedure and review». In: *Animal Behaviour* 55.4, pp. 827–843.

Dempster, A.P., N.M. Laird e D.B. Rubin (1977). «Maximum likelihood from incomplete data via the EM algorithm». In: *Journal of the royal statistical society. Series B (methodological)* 39.1, pp. 1–38.

Erdös, P. e A. Rényi (1959). «On random graphs, I». In: *Publicationes Mathematicae (Debrecen)* 6, pp. 290–297.

— (1960). «On the evolution of random graphs». In: *Publ. Math. Inst. Hung. Acad. Sci* 5.1, pp. 17–60.

Fawcett, T. (2006). «An introduction to ROC analysis». In: *Pattern recognition letters* 27.8, pp. 861–874.

Fienberg, S.E. e S.S. Wasserman (1981). «Categorical data analysis of single sociometric relations». In: *Sociological methodology* 12, pp. 156–192.

Fruchterman, T.M.J. e E.M. Reingold (1991). «Graph drawing by force-directed placement». In: *Software: Practice and experience* 21.11, pp. 1129–1164.

Geman, S. e D. Geman (1984). «Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images». In: *IEEE Transactions on pattern analysis and machine intelligence* 6, pp. 721–741.

Hand, D.J. e R.J. Till (2001). «A simple generalisation of the area under the ROC curve for multiple class classification problems». In: *Machine learning* 45.2, pp. 171–186.

Handcock, M.S., A.E. Raftery e J.M. Tantrum (2007). «Model-based clustering for social networks». In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 170.2, pp. 301–354.

Hanley, J.A. e B.J. McNeil (1982). «The meaning and use of the area under a receiver operating characteristic (ROC) curve». In: *Radiology* 143.1, pp. 29–36.

Hastings, W.K. (1970). «Monte Carlo sampling methods using Markov chains and their applications». In: *Biometrika* 57.1, pp. 97–109.

Hoff, P.D., A.E. Raftery e M.S. Handcock (2002). «Latent space approaches to social network analysis». In: *Journal of the american Statistical association* 97.460, pp. 1090–1098.

Holland, P.W., K.B. Laskey e S. Leinhardt (1983). «Stochastic blockmodels: First steps». In: *Social networks* 5.2, pp. 109–137.

Karrer, B. e M.E.J. Newman (2011). «Stochastic blockmodels and community structure in networks». In: *Physical Review E* 83.1. URL: https://doi.org/10.1103/PhysRevE.83.016107.

Kirkpatrick, S., C.D. Gelatt e M.P. Vecchi (1983). «Optimization by simulated annealing». In: *Science* 220.4598, pp. 671–680.

Kobourov, S.G. (2013). «Force-Directed Drawing Algorithms». In: *Handbook of Graph Drawing and Visualization*. A cura di Roberto Tamassia. CRC Press. Cap. 12, pp. 383–408.

Kolaczyk, E.D. (2009). *Statistical Analysis of Network Data. Methods and Models*. Springer.

Rapporto annuale Istat 2015. *La situazione del Paese*. Ultima visita il 08/02/2017. URL: http://www.istat.it/it/archivio/159350/.

Metropolis, N. et al. (1953). «Equation of state calculations by fast computing machines». In: *The journal of chemical physics* 21.6, pp. 1087–1092.

Nowicki, K. e T.A.B. Snijders (2001). «Estimation and prediction for stochastic blockstructures». In: *Journal of the American Statistical Association* 96.455, pp. 1077–1087.

Paul, S. e Y. Chen (2015). «Community detection in multi-relational data with restricted multi-layer stochastic blockmodel». In: *arXiv preprint*. URL: arXiv:1506.02699.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Rossum, G. (1995). *Python Reference Manual*. Amsterdam, The Netherlands: CWI (Centre for Mathematics e Computer Science). URL: https://docs.python.org/2.0/ref/ref.html.

Schwarz, G. et al. (1978). «Estimating the dimension of a model». In: *The annals of statistics* 6.2, pp. 461–464.

Snijders, T.A.B. e K. Nowicki (1997). «Estimation and prediction for stochastic blockmodels for graphs with latent block structure». In: *Journal of classification* 14.1, pp. 75–100.

VQR 2011-2014. *Valutazione della Qualità della Ricerca 2011-2014 ANVUR (Agenzia Nazionale di Valutazione del sistema Universitario e della Ricerca)*. Ultima visita il 24/02/2017. URL: http://www.anvur.org/rapporto-2016/.

Way, S.F., D.B. Larremore e A. Clauset (2016). «Gender, Productivity, and Prestige in Computer Science Faculty Hiring Networks». In: *arXiv preprint*. URL: arXiv:1602.00795.

Ringraziamenti

A Caterina, per il suo enorme e prezioso aiuto e lavoro a suon di mail a distanza con otto ore di differenza.

Ai miei genitori, è grazie a voi se sono diventata la persona che sono. Il vostro appoggio è sempre stato fondamentale in ogni scelta.

A mia nonna, se sono il tuo bastone della vecchiaia, tu sei il mio sostegno da quando sono nata.

Ai miei cugini-fratelli, un pezzo di cuore, anche quando non riesco a esservi vicina come vorrei. A tutta la famiglia che in ogni modo possibile mi supporta e sopporta.

A Elisa e Giulia, presenti anche nei periodi più pieni e difficili, so che sarete sempre al mio fianco e vi sono grata di ogni messaggio e ogni risata e ogni sushi.

A Silvia, Giulia, Chiara, Marco e Santiago, per ogni serata e ogni litigio post-risiko e per i viaggi fatti e quelli che faremo insieme.

A Chiara, per essere diventata nell'ultimo anno quello che speravo da sempre diventassi (e per essere la mia grafica di fiducia).

Infine, ad Alessandro, senza il quale tutto sarebbe meno colorato. Grazie per la nostra vita, appena iniziata.