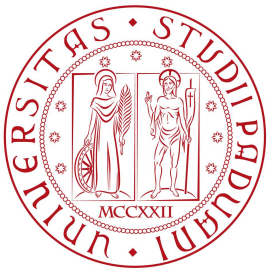
Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in

Scienze Statistiche



Analisi delle reti di partecipazione competitiva nel Powerlifting italiano: un'applicazione al dataset OpenPowerlifting

Relatrice: Prof.ssa Mariangela Guidolin
Dipartimento di Scienze Statistiche

Laureanda: Alice Cappella

Matricola n. 2095652

Anno Accademico 2023/2024

Alla mia famiglia

Indice

Introduzione

Viviamo in un mondo profondamente connesso. Questa affermazione riflette una realtà sempre più evidente, che giustifica la crescente diffusione dell'approccio basato sui modelli di rete. Quest'ultimo si presta a utilizzi sempre più ampi, anche in ambiti apparentemente distanti.

Una rete può essere descritta come un insieme di persone o oggetti connessi tra loro da un qualche tipo di relazione. Comprendere le dinamiche delle reti può fornire informazioni preziose sulle interazioni che si verificano tra quelli che vengono comunemente definiti nodi.

In questo elaborato viene proposta un'applicazione delle tecniche di analisi dei dati di rete nell'ambito sportivo italiano, nello specifico al *Powerlifting*, uno sport di forza che comprende tre alzate: *squat*, distensioni su panca e stacco da terra. A tal fine, viene utilizzato il *dataset OpenPowerlifting*, un insieme di dati *open source* relativo alle prestazioni degli atleti nelle competizioni a cui hanno preso parte. In questo contesto, la rete è rappresentata dagli atleti legati tra loro attraverso la partecipazione alla medesima competizione.

Esplorare queste relazioni consente di individuare atleti che, avendo partecipato a molte competizioni, presentano un numero elevato di connessioni. Questi individui, avendo accumulato una maggiore esperienza, possono risultare particolarmente competitivi. L'osservazione delle loro strategie può fornire spunti utili per gli altri atleti e i loro allenatori, contribuendo a definire programmazioni di allenamento più efficaci. Inoltre, la disponibilità di dati in diversi istanti temporali consente di valutare come le partecipazioni competitive e le relazioni tra gli atleti cambino nel tempo.

La tesi si sviluppa in quattro capitoli. Il Capitolo 1 è dedicato a una panoramica generale sul *Powerlifting* e sul regolamento vigente alle competizioni. Verrà poi descritto il *dataset* rilasciato da *OpenPowerlifting* e, infine, presentata una breve analisi esplorativa di alcune variabili rilevanti, con l'obiettivo di approfondire la comprensione di questo sport, delle caratteristiche degli atleti e delle competizioni.

Nel Capitolo 2 verranno approfonditi i concetti base dell'analisi dei dati di rete. Dopo aver fornito le definizioni chiave e le statistiche descrittive più comunemente utilizzate, si discuteranno le proprietà tipiche delle reti reali. L'attenzione sarà poi focalizzata su due particolari tipologie di reti: quelle bipartite e quelle dinamiche. Il capitolo si concluderà con l'applicazione delle tecniche di analisi delle reti ai dati in esame, analizzando come queste possano fornire una visione più chiara delle dinamiche di partecipazione nel contesto del *Powerlifting*.

Nel Capitolo 3 si esamineranno le tecniche di visualizzazione grafica, con particolare riferimento agli algoritmi di *force-directed placements*, utilizzati per ottimizzare la disposizione dei nodi. Si studierà poi il rilevamento delle comunità, che mira ad identificare gruppi di nodi con una forte connessione interna e una minore interazione esterna. Si conclude con l'utilizzo delle tecniche viste per il *dataset OpenPowerlifting*.

Infine, nel Capitolo 4 vengono presentati diversi modelli per l'analisi delle reti. Si inizierà con il modello basato sulla decomposizione ANOVA e il *Social Relation Model*, seguiti dal *Social Relation Regression Model*, per poi passare all'*Additive and Multiplicative Effects Model* (AME). Successivamente, si approfondiranno i modelli di trasformazione per reti non gaussiane, una generalizzazione dei modelli AME in presenza di dati binari o ordinali. Il capitolo si concluderà con l'applicazione di questi modelli ai dati di partecipazione competitiva nel *Powerlifting*.

Capitolo 1

Il *Powerlifting*

Il *Powerlifting* consiste in uno sport di forza che coinvolge tre movimenti: lo *squat*, le distensioni su panca e lo stacco da terra. Questo sport, similmente al più noto *Weightlifting*, prevede tre tentativi per ogni movimento, in cui l'atleta cerca di sollevare il maggior peso caricato tramite dischi su una lunga barra in acciaio, chiamata bilanciere. La differenza con il *Weightlifting* risiede principalmente nella tipologia di sollevamenti. Nel *Powerlifting*, infatti, il *range* dei movimenti, vale a dire la loro ampiezza, risulta ridotto consentendo all'atleta di sollevare un peso maggiore.

1.1 Regolamento

In questa sezione si descriverà il regolamento che disciplina le gare di *Powerlifting*. In particolare, verrà seguito lo standard definito dall'IPF, adottato ad ogni gara di tutti i livelli (IPF, 2023).

Le competizioni sono composte da tre prove di sollevamento: *squat*, distensioni su panca e stacco da terra (eseguite con questo ordine). Ogni atleta ha a disposizione tre tentativi per ogni sollevamento e il totale viene calcolato tenendo in considerazione la migliore alzata eseguita. Inoltre, se due atleti registrano lo stesso totale sollevato, l'atleta con peso corporeo minore sarà classificato prima di quello con peso maggiore.

Ogni sollevamento viene sottoposto alla valutazione di tre giudici, che determinano se l'alzata risulta valida (luce bianca) o nulla (luce rossa). L'alzata si ritiene valida se l'atleta ottiene almeno due luci bianche. Qualora uno o più giudici ritenessero nullo un tentativo di sollevamento, dopo che la luce rossa viene attivata, alzerà/alzeranno un cartello per indicare la motivazione della valutazione effettuata.

1.1.1 Le alzate

Nello *squat*, l'atleta, dopo essersi posizionato il bilanciere sulle spalle, impugnandolo con le mani, e assunto una posizione eretta con ginocchia distese, deve piegarle e abbassare il corpo, come per sedersi, per poi ritornare alla posizione di partenza.

A seguito dell'annuncio che informa che il bilanciere è stato caricato, ogni atleta ha a disposizione un minuto per sganciare il bilanciere dal *rack*[1], superato questo tempo il tentativo verrà giudicato nullo. Vengono annunciati anche l'inizio, per la fase di discesa, e la fine dell'alzata, vale a dire il riposizionamento del bilanciere sul *rack*. Se l'atleta non rispetta questi segnali la prova sarà nulla.

Figura 1.1: *Squat*: validità dell'alzata.

[1]Struttura metallica regolabile per sostenere il bilanciere prima e dopo l'esecuzione delle alzate.

Un altro parametro da rispettare per far sì che una prova di sollevamento risulti valida è la profondità raggiunta: l'atleta deve "rompere il parallelo", vale a dire che, nella porzione di movimento in cui si trova in basso, l'articolazione dell'anca si deve trovare al di sotto di quella del ginocchio. Questa è la causa più comune delle prove nulle e viene indicata, oltre alla luce rossa, con un cartellino rosso.

Nelle distensioni, l'atleta si trova disteso sulla panca con testa e spalle a contatto con essa e piedi saldamente a terra. Le mani devono impugnare il bilanciere ad una distanza di 81 cm, i cui limiti sono segnati sul bilanciere stesso. Una volta che il bilanciere viene rimosso dal *rack*, l'atleta, piegando i gomiti, deve abbassare il bilanciere fino a che questo non tocchi un qualsiasi punto che va dal torace all'area addominale, ma al di sopra della cintura, per poi tornare alla posizione iniziale.

Figura 1.2: Distensioni su panca: validità dell'alzata.

Per questa alzata, oltre all'inizio e alla fine, viene annunciato anche quando l'atleta può, dalla fase in cui il bilanciere è a contatto con il corpo, ritornare alla posizione di partenza. Anche in questo caso, così come nello *squat*, dopo il segnale per l'inizio, l'atleta ha a disposizione un minuto di tempo per iniziare il sollevamento.

La prova sarà nulla anche se il bilanciere non entra in contatto con il corpo oppure se durante l'alzata l'atleta cambia la posizione di testa, spalle, glutei o delle mani sul bilanciere. Inoltre, affinché il tentativo venga considerato valido, l'aspetto più importante rimane la profondità raggiunta: le articolazioni di entrambi i gomiti devono raggiungere un punto al di sotto delle articolazioni delle spalle.

Infine, lo stacco da terra, così come suggerisce il nome del movimento, prevede che l'atleta sollevi il bilanciere da terra fino a raggiungere una posizione eretta con ginocchia stese e spalle proiettate all'indietro, per poi riposizionare il bilanciere nella posizione di partenza. Non ci sono regole né sulla presa delle mani né sulla distanza tra i piedi. Per quanto riguarda la presa del bilanciere, le tipologie più utilizzate sono la prona, ossia con i pollici rivolti verso l'interno, e la mista, ovvero con i palmi in direzioni differenti.

In base alla distanza tra i piedi, invece, si possono avere due tipi di stacco da terra:

- *sumo*, in cui i piedi sono molto distanti e si prende il bilanciere internamente rispetto alle gambe;
- *regular* o *conventional*, dove i piedi sono più ravvicinati (posizionati sotto i fianchi) e l'atleta prende il bilanciere all'esterno rispetto alle gambe.

A differenza delle altre alzate, non viene dato un segnale per l'inizio dell'alzata ma l'atleta ha sempre a disposizione un minuto da quando viene annunciato il bilanciere pronto per iniziare il sollevamento. Dopo aver sollevato il bilanciere, l'atleta dovrà attendere il segnale per poterlo abbassare. Se l'atleta scende prima di quando è consentito o perde la presa, la prova sarà nulla.

Le due regole principali riguardano le ginocchia, che devono risultare stese, e la posizione delle spalle, che dovranno essere rivolte all'indietro. Non è poi possibile "infilare" il bilanciere, ossia appoggiare il bilanciere sulle cosce per cercare di terminare il movimento.

Figura 1.3: Stacco da terra (*sumo* con presa prona): esecuzione e validità dell'alzata.

In generale, le prove saranno nulle se il sollevamento non viene completato. Per approfondire tutte le motivazioni di prove nulle si rimanda a IPF (2023).

1.1.2 Categorie ed equipaggiamenti

Gli atleti vengono suddivisi in base al sesso, all'età e al peso. Nello specifico, le categorie di età sono:

- *open*: dai 14 anni in su;
- *sub-junior*: dai 14 ai 18 anni;
- *junior*: dai 19 ai 23 anni;
- *senior*: dai 24 ai 39 anni;
- *master* I: dai 40 ai 49 anni;
- *master* II: dai 50 ai 59 anni;
- *master* III: dai 60 ai 69 anni;
- *master* IV: dai 70 anni in su.

Le categorie di peso, suddivise in base al sesso, vengono riportate nella Tabella 1.1.

Uomini		**Donne**	
Categoria	Peso (kg)	Categoria	Peso (kg)
53kg	(0,53]	43kg	(0,43]
59kg	(0,59]	47kg	(0,47]
66kg	(59,66]	52kg	(47,52]
74kg	(66,74]	57kg	(52,57]
83kg	(74,83]	63kg	(57,63]
93kg	(83,93]	69kg	(63,69]
105kg	(93,105]	76kg	(69,76]
120kg	(105,120]	84kg	(76,84]
120+kg	(120,+∞)	84+kg	(84,+∞)

Tabella 1.1: Categoria di peso per uomini e donne.

Si specifica che le categorie 53kg per gli uomini e 43kg per le donne sono valide solamente per atleti *sub-junior* e *junior*.

In base all'attrezzatura personale concessa si parla di competizioni attrezzate o *equipped* e competizioni *raw* o *classic*. Nelle competizioni attrezzate, l'atleta può indossare il cosiddetto corpetto di supporto mentre in quelle *raw* vengono indossati solamente corpetti non di supporto. I corpetti utilizzati nelle competizioni *raw* possono essere indossati anche in quelle attrezzate, ma non vale il viceversa. Altri equipaggiamenti consentiti sono la cintura, le ginocchiere e i polsini. Tutta l'attrezzatura personale deve essere autorizzata ed essere presente nella "Lista approvata di abbigliamento e attrezzatura per gare IPF". Per questo motivo, ad ogni competizione, viene effettuato il controllo dell'attrezzatura personale.

1.2 Il *dataset*

In questo elaborato verranno utilizzati i dati rilasciati da *OpenPowerlifting*, "un progetto di servizio alla comunità per creare un archivio permanente e aperto dei dati mondiali sul *Powerlifting*" (OpenPowerlifting, 2024). È quindi possibile avere accesso diretto a questi dati, senza la necessità di effettuare *web scraping*[2].

Il *dataset* include informazioni sulle competizioni di atleti di *Powerlifting* a livello globale. In particolare, ogni riga è relativa alle prestazioni e alle caratteristiche di uno specifico atleta in una determinata competizione. Le dimensioni sono considerevoli, con una numerosità di oltre tre milioni. A fronte di questo, per facilitare la gestione dei dati e garantire una maggiore interpretazione, l'analisi svolta si limita a considerare le gare svolte in Italia.

1.2.1 Le variabili

In questo paragrafo verranno presentate le variabili più rilevanti tra le 42 contenute nell'insieme di dati in esame, con descrizioni basate sulla documentazione disponibile in OpenPowerlifting Data Service (2024). Per maggiori dettagli sulle altre variabili, sui punteggi utilizzati per il confronto delle *performance* e sulle analisi preliminari svolte, si veda A.2.

- `Name`: il nome dell'atleta.
- `Sex`: la categoria di sesso in cui l'atleta compete (`M`, `F`, `Mx`, dove `Mx` è una categoria sessuale generica adatta ad atleti non binari).
- `Age`: l'età dell'atleta alla data di inizio del torneo (se nota). Può essere di due tipologie:
  - esatta: numero intero;
  - approssimata: $n + 0.5$ per cui la possibile età sarà n o $n + 1$.

[2]Il *web scraping* è una procedura informatica che consente di raccogliere dati da un sito *web*.

- `Division`: la divisione di età della competizione.
- `BodyweightKg`: il peso corporeo dell'atleta registrato alla competizione, arrotondato a due decimali.
- `Best3SquatKg`, `Best3BenchKg`, `Best3DeadliftKg`: il massimo dei tre tentativi riusciti per l'alzata. Alcune federazioni non riportano quanto è stato sollevato nelle singole prove ma solo il valore di queste variabili. I valori negativi vengono, raramente, utilizzati per segnalare il peso più basso che l'atleta ha tentato e non ha superato.
- `TotalKg`: la somma delle variabili `Best3SquatKg`, `Best3BenchKg`, `Best3DeadliftKg`. Se si ha una prova nulla in almeno un'alzata o l'atleta è stato squalificato per qualche ragione, `TotalKg` è vuoto. Raramente viene riportato solo il totale ma nessun dato sulle singole alzate.
- `Event`:
  - `SBD`: *Squat-Bench-Deadlift* ("*Full Power*");
  - `BD`, `SD` e `SB`: rispettivamente *Bench-Deadlift* ("*Ironman*" o "*Push-Pull*"), *Squat-Deadlift* e *Squat-Bench*;
  - `S`, `B` e `D`: rispettivamente, solo *Squat*, solo *Bench* e solo *Deadlift*.
- `Equipment`: la categoria di attrezzatura sotto la quale sono stati eseguiti gli esercizi.
  - `Raw`: *bare knees* o *knee sleeves* (senza o con ginocchiere);
  - `Wraps`: *knee wraps* (fasce per le ginocchia);
  - `Single-ply`: *single-ply suits*;
  - `Multi-ply`: *multi-ply suits*;
  - `Unlimited`: *multi-ply suits* o *rubberized gear*;
  - `Straps`: *straps* (fascette, facilitano la presa) per lo stacco da terra. Utilizzate soprattutto per esibizioni e non competizioni reali.

- `Dots`: un numero positivo se i punti DOTS possono essere calcolati, vuoto se l'atleta è stato squalificato.
- `Date`: la data di inizio dell'evento (YYYY-MM-DD).
- `MeetState`: lo Stato/Provincia o Regione in cui si è svolta la competizione.
- `MeetName`: il nome della competizione.

1.3 Analisi esplorativa

Di seguito viene effettuata una breve analisi esplorativa del *dataset*, attraverso una serie di grafici univariati e bivariati. Si procederà inizialmente con l'analisi delle variabili legate alle caratteristiche e alle *performance* degli atleti, per poi concentrarsi maggiormente su quelle relative alle competizioni.

A seguito delle operazioni preliminari, descritte in A.2.3, l'insieme di dati conta 30615 osservazioni, di cui 23221 (76%) riguardano atleti di sesso maschile e 7394 (24%) di sesso femminile. Questo risultato non sorprende, poiché il *Powerlifting* ha origine come sport prevalentemente maschile e, di conseguenza, ha avuto uno sviluppo più marcato tra gli uomini. Sebbene negli ultimi anni questo sport si stia diffondendo anche tra le donne, la predominanza maschile persiste.

I grafici in Figura 1.4 mostrano gli istogrammi della variabile relativa al peso corporeo. La distribuzione del peso per le atlete appare asimmetrica, con una maggiore frequenza di valori al di sotto degli 80kg e una coda destra più lunga. Per quanto riguarda gli atleti, la distribuzione risulta invece più simmetrica. La maggior parte dei valori si trovano tra i 70kg e i 100kg ma vi è una buona percentuale anche di soggetti con un peso corporeo superiore. Si specifica che sono stati esclusi 31 atleti con peso al di fuori del *range* rappresentato allo scopo di migliorare la visualizzazione della distribuzione.

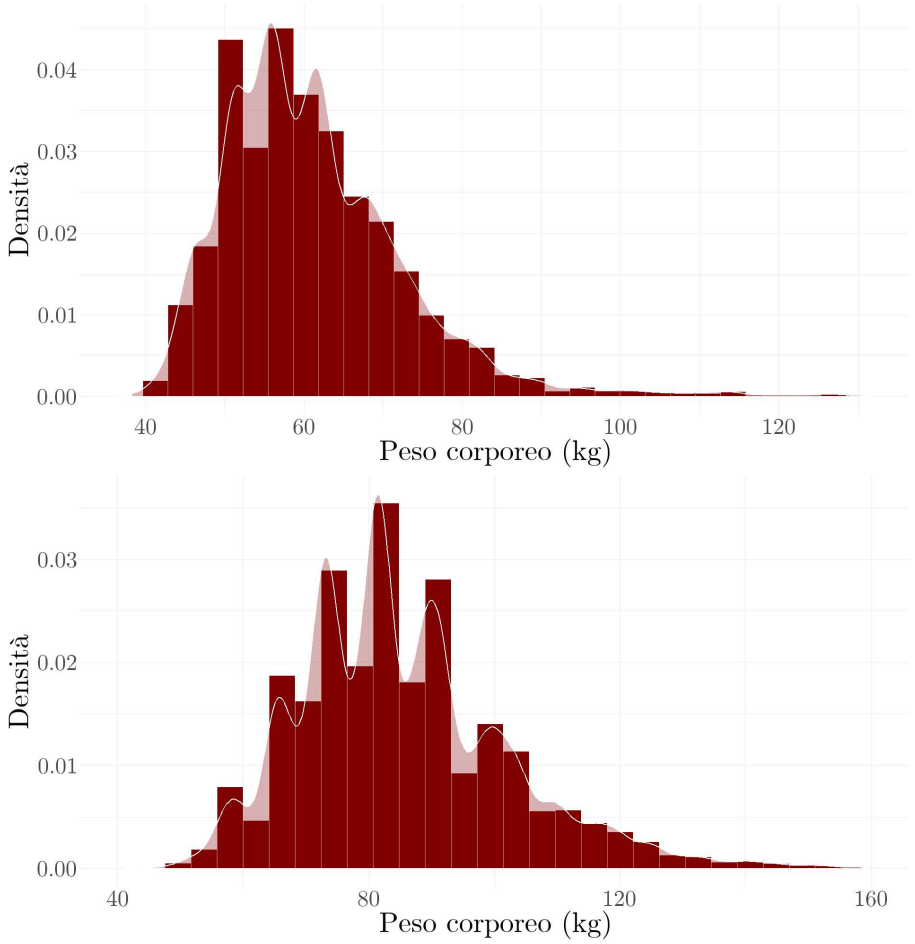


Figura 1.4: Istogramma, con sovrapposta densità, del peso corporeo per donne (in alto) e uomini (in basso).

Si continua a considerare il peso corporeo mettendolo, tuttavia, in relazione al totale sollevato (variabile `TotalKg`). Si ricorda che nel caso di atleti squalificati, indicati con punti rossi nelle rappresentazioni in Figura 1.5, il totale sollevato non viene calcolato e risulta quindi nullo. In entrambi i grafici si osserva una relazione positiva tra il peso corporeo e il totale sollevato. Come ci si attende, i valori del totale sollevato dagli atleti maschi sono superiori rispetto a quelli delle donne. Inoltre, in entrambi i casi si notano due gruppi sufficientemente distinti. Questa distinzione è dovuta alla tipologia di competizione svolta. Il gruppo inferiore, con totale sollevato più ridotto, è relativo a gare che vedono coinvolte una o al massimo due alzate, mentre il gruppo superiore è relativo a competizioni che prevedevano tre alzate.

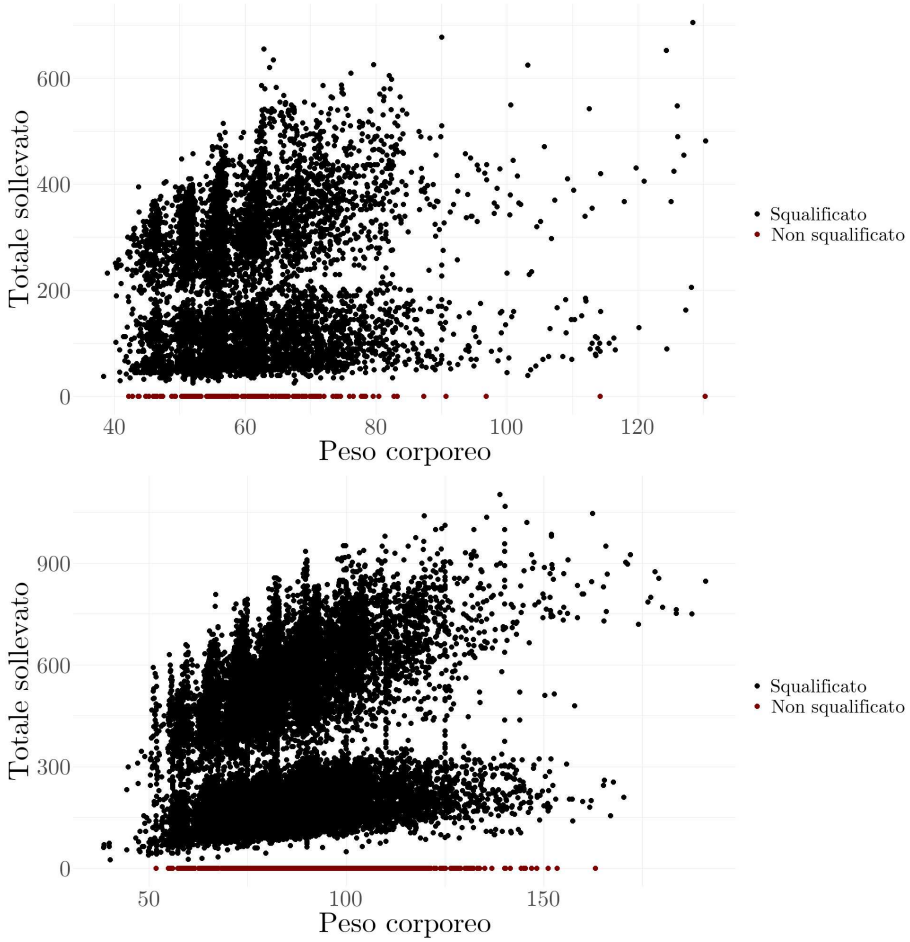


Figura 1.5: Grafico di dispersione del totale sollevato rispetto al peso corporeo per donne (in alto) e uomini (in basso)

Si prosegue ora a considerare le variabili relative ai migliori tentativi nelle tre singole alzate. Nei grafici rappresentati in Figura 1.6 si osservano densità bimodali, data la presenza di tentativi falliti indicati da 0. Sia per gli uomini che per le donne si nota una maggiore frequenza di tentativi falliti per l'alzata dello *squat*. Le motivazioni possono essere molteplici. In primo luogo, in questa alzata è comune che gli arbitri giudichino la prova come nulla. Durante l'esecuzione dello *squat*, infatti, è difficile per l'atleta determinare quando viene passato il parallelo. Una seconda motivazione potrebbe essere riconducibile all'ordine delle alzate. Essendo il primo sollevamento, gli allenatori potrebbero puntare maggiormente in questa alzata per una maggiore freschezza fisica dell'atleta. È comprensibile che, con il

progredire della gara, i preparatori degli atleti adottino un approccio più conservativo, specialmente nello stacco da terra, poiché ottenere esclusivamente tentativi falliti in un'alzata comporterebbe la squalifica dell'atleta, compromettendo il risultato finale della competizione. Infine, trovandosi all'inizio della gara, sicuramente l'aspetto di pressione psicologica (personale e degli avversari) gioca un ruolo importante e potrebbe influire sulla riuscita dell'alzata.

Trascurando i tentativi falliti, l'alzata che mostra un peso medio sollevato maggiore è lo stacco da terra, seguita dallo *squat* e dalle distensioni su panca. L'ultima alzata coinvolge infatti muscoli più piccoli rispetto alle altre due, portando a sollevare un quantitativo più ridotto.

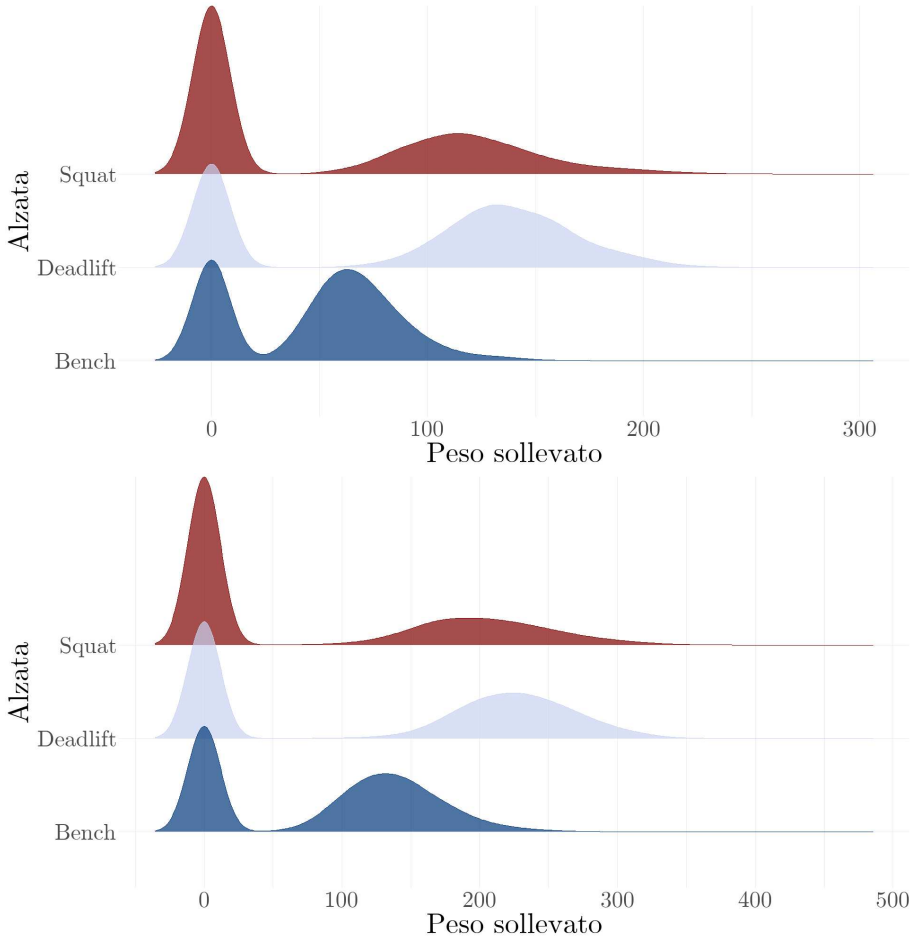


Figura 1.6: Densità del peso sollevato al miglior tentativo nelle singole alzate da donne (in alto) e uomini (in basso).

Dal grafico relativo alla variabile `Division`, riportato in Figura 1.7, si nota come la categoria di età più frequente sia *open*, che comprende atleti dai 14 anni in su, seguita da *junior* e *senior*. Dall'ultima barra, è chiaro che risulti molto raro che in una competizione ci sia un ospite (0.01%). Inoltre, per lo 0,43% degli atleti, per i quali non è stata indicata la divisione, non è stato possibile procedere con l'inputazione a causa della mancanza del dato relativo all'età. Questi rientrano quindi nella categoria indicata come "*missing*".

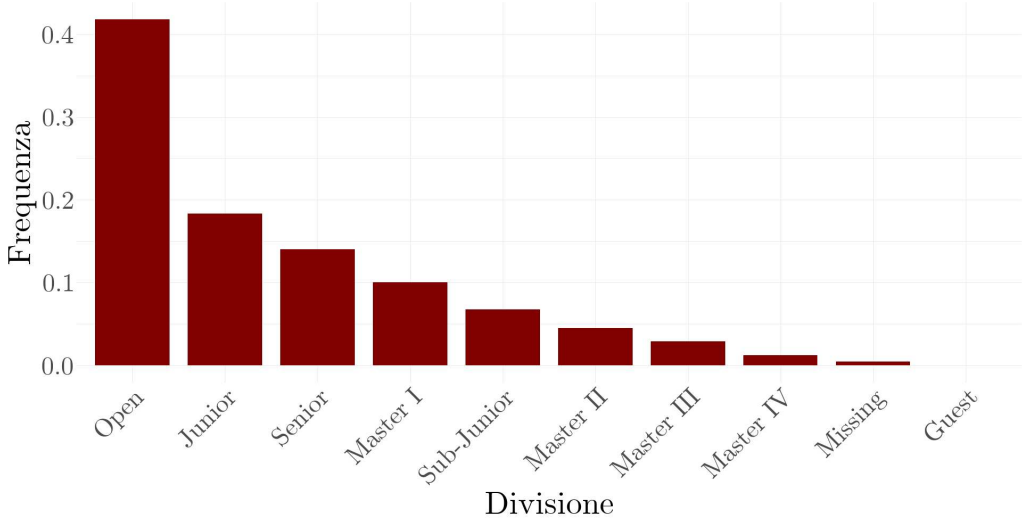


Figura 1.7: Grafico a barre della divisione di età.

Proseguendo con l'analisi dell'informazione fornita dalla divisione degli atleti è possibile valutare in corrispondenza di quale categoria di età si osservano i valori più elevati dei punti DOTS (Figura 1.8).

Considerando la mediana per divisione, ed escludendo per il momento gli atleti ospiti delle competizioni, il valore maggiore si ottiene in corrispondenza della categoria *junior*, che includendo atleti molto giovani risulta molto competitiva. La mediana va poi a diminuire all'aumentare dell'età degli atleti. I pochi atleti che sono stati invitati come ospiti alle competizioni sono molto forti e questo è evidenziato da punteggi DOTS molto elevati.

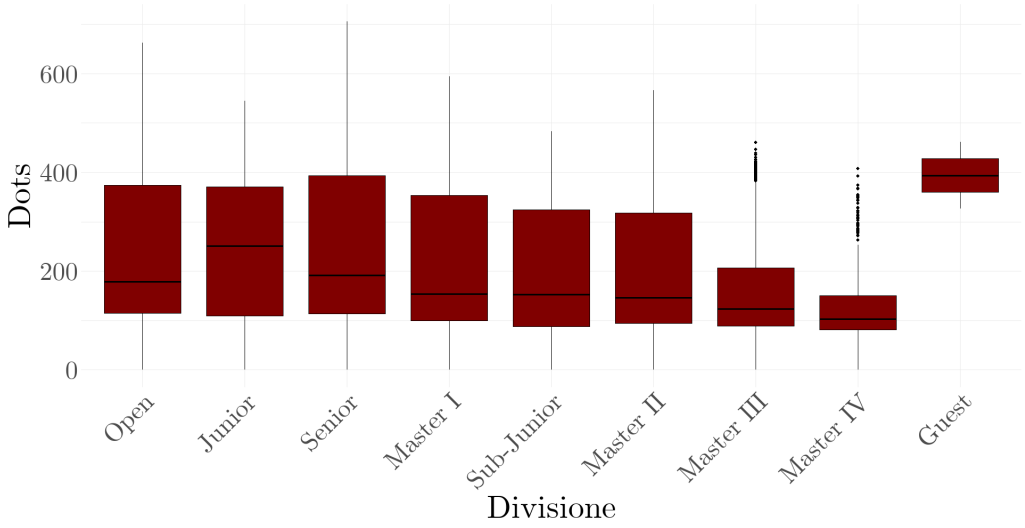


Figura 1.8: Boxplot dei punti DOTS in funzione della divisione.

L'attenzione si concentra ora sulle variabili connesse agli eventi. Come ci si poteva attendere, dal grafico in Figura 1.9, risulta evidente che le competizioni SBD sono le più frequenti, seguite da quelle con una singola alzata, vedendo un numero maggiore di gare che coinvolgono le distensioni su panca o lo stacco da terra, e da quelle con due alzate.

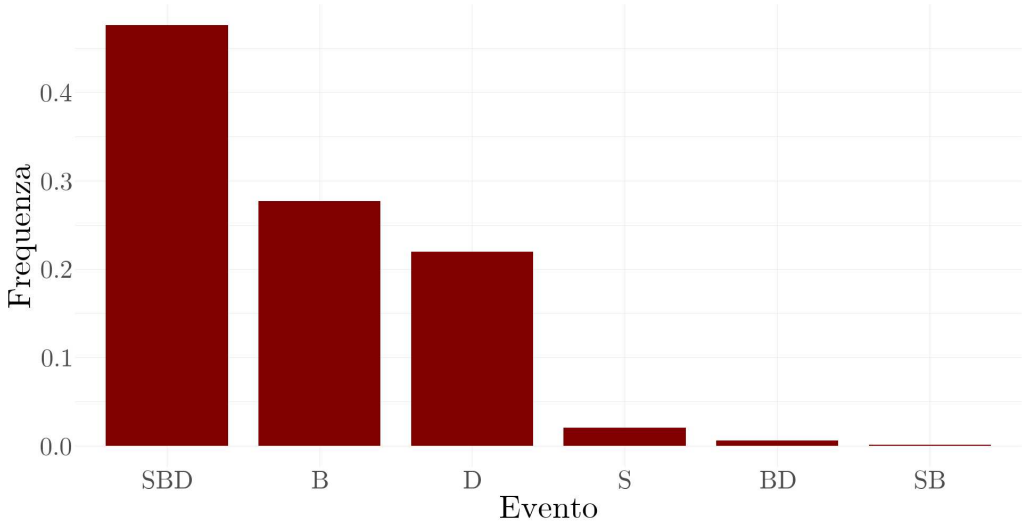


Figura 1.9: Grafico a barre della tipologia di competizione (`Event`).

La variabile `MeetName`, relativa al nome delle competizioni, riveste un'importanza fondamentale. Consentirà, infatti, di utilizzare i dati a disposizione con un'ottica di rete, considerando le comuni partecipazioni competitive come connessione tra gli atleti. In totale si registrano 167 competizioni svolte tra il 1979 e il 2024.

La serie storica del numero di competizioni, illustrata nella parte superiore della Figura 1.10, mostra, dopo un'iniziale crescita, un periodo di stabilità dal 1984 al 2000. Durante questo decennio, il numero di eventi si mantiene su valori relativamente contenuti, con una sola competizione svolta ogni anno. Successivamente, a partire dal 2000, si osserva un'inversione di tendenza con un aumento significativo del numero di competizioni. Questo incremento potrebbe essere attribuito alla nascita e al successivo sviluppo della *Federazione Italiana PowerLifting* (FIPL). La crescita si intensifica ulteriormente dal 2013 raggiungendo un massimo di 16 competizioni svolte nel 2019. Tuttavia, la pandemia di COVID-19 nel 2020 ha portato a un forte calo, interrompendo questa tendenza positiva. Un aumento delle competizioni è stato poi registrato dal 2020 al 2023.

L'andamento del numero di atleti, rappresentato nel grafico in basso della Figura 1.10, appare più regolare rispetto a quello del numero di competizioni appena descritto. Anche in questo caso, fino al 2000, si osserva una stabilità nei partecipanti. A partire dal 2000, si registra un incremento progressivo, indicando uno sviluppo del *Powerlifting* in Italia, con nuovi atleti che iniziano a competere ogni anno. La crescita del numero di atleti risulta particolarmente evidente a partire dal 2010, coincidente con l'aumento delle competizioni svolte. Così come per le competizioni, il coinvolgimento degli atleti ha subito un brusco calo dopo il picco del 2019, a causa della pandemia di COVID-19.

È importante specificare che l'analisi non considera l'ultimo anno, poiché i dati sono stati raccolti all'inizio del 2024 e il numero di eventi era ancora limitato.

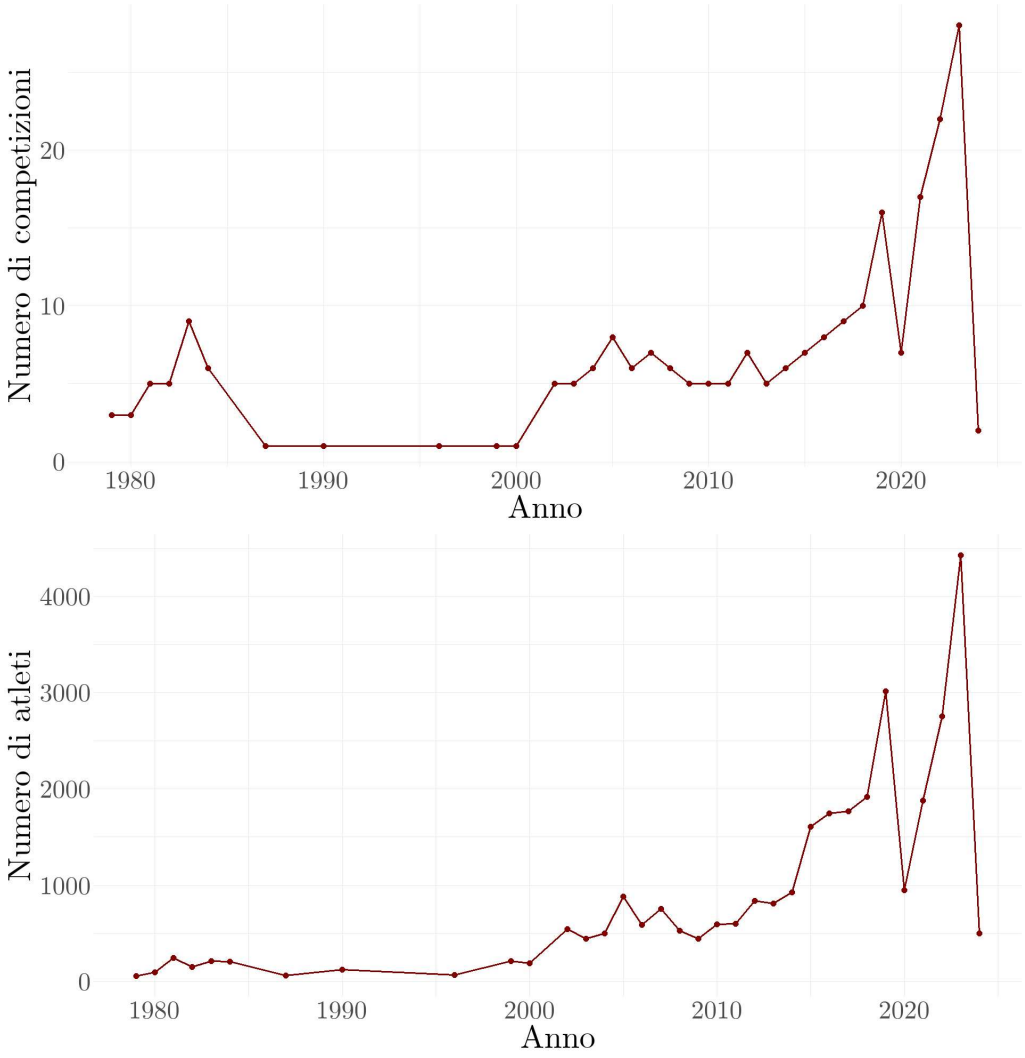


Figura 1.10: Serie storica del numero di competizioni (in alto) e di atleti (in basso).

La cartina geografica mostrata nella Figura 1.11 illustra le sedi delle competizioni nel corso degli anni (punti bianchi), con colore delle aree proporzionale al numero di gare ospitate. In grigio vengono indicate le Regioni nelle quali non si è mai tenuta una competizione nel periodo considerato. Si osserva che la Lombardia è la regione con il maggior numero di competizioni. Si precisa, inoltre, che sono state esclusi 14 eventi per i quali non è stata fornita la città di svolgimento.

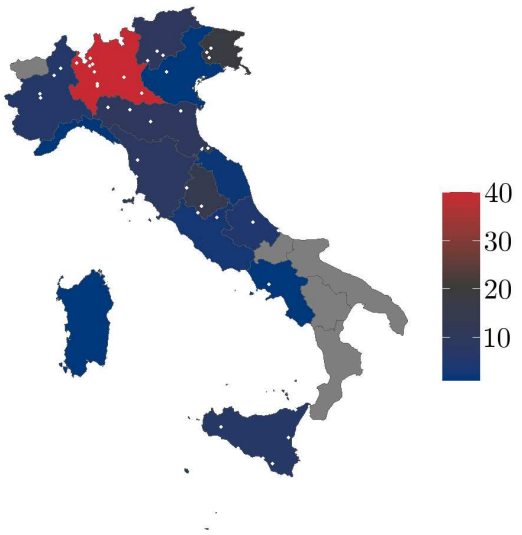


Figura 1.11: Numero di competizioni per Regione italiana.

È infine possibile analizzare la correlazione tra le variabili di natura numerica (Figura 1.12).

Come evidenziato nell'analisi del grafico in Figura 1.5, il peso corporeo è correlato positivamente con il totale sollevato.

I punti DOTS sono correlati positivamente con il totale sollevato, essendo una quantità coinvolta nella formula per il suo calcolo al numeratore.

Ponendo l'attenzione sui tre tentativi per ogni alzata, si evidenzia in tutti e tre i casi una correlazione positiva tra le prime due prove. Inoltre, nello stacco da terra si osserva una leggera correlazione negativa tra i primi due tentativi e l'ultimo. Essendo l'ultima prova della gara, si può supporre che la fatica accumulata nelle prime due alzate porti ad un numero maggiore di tentativi falliti (indicati con zero). Allo stesso tempo, l'ultima alzata è spesso utilizzata in modo strategico dagli atleti che hanno una buona posizione in classifica, sia per cercare di ribaltare il risultato, sia per stabilire nuovi *record* personali. In entrambi i casi, un aumento eccessivo del carico utilizzato potrebbe portare a una maggiore probabilità di tentativi falliti.

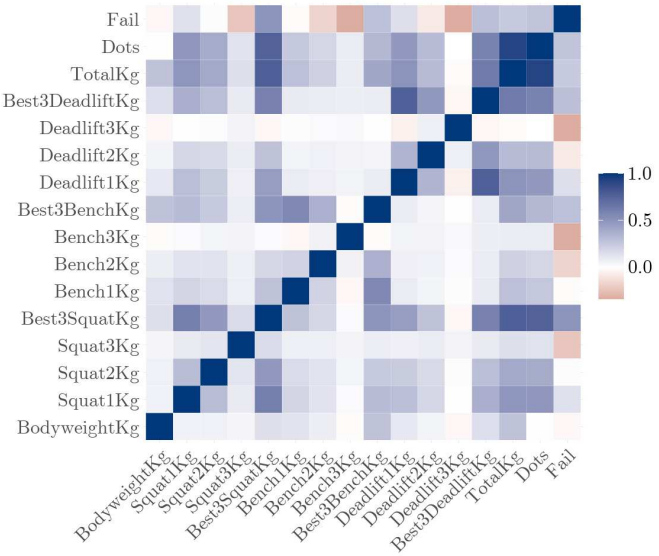


Figura 1.12: Matrice di correlazione delle variabili numeriche.

Capitolo 2

Dati di rete

Sebbene il termine "rete sociale" richiami il mondo dei *social media*, nella teoria matematica si riferisce ad una qualsiasi rete di persone o entità, collegate tra loro tramite una qualche forma di connessione.

I primi studi di reti sociali si riconducono agli anni '30, quando Jacobs Moreno, uno psicologo tedesco, analizzò la dinamica delle relazioni sociali tra bambini in una classe di scuola elementare. Moreno ideò in questo modo il cosiddetto sociogramma, oggi noto come grafo, che consente la rappresentazione delle connessioni tra persone mediante punti collegati da linee.

Già a partire da questo esempio, agli albori della teoria delle reti sociali, e dalla generalità della definizione di rete si può intuire come, al giorno d'oggi, le applicazioni in questo ambito siano molto diffuse e coinvolgano diversi settori, come la fisica, la biologia, l'informatica, l'economia, la sociologia e la psicologia, solo per citarne alcuni.

In questo capitolo verrà fornita una panoramica generale della teoria delle reti, con particolare attenzione alla definizione dei termini e alle statistiche descrittive più rilevanti. Dopo aver esaminato le proprietà comunemente riscontrate nelle reti e introdotto due tipologie specifiche, si passerà all'applicazione delle nozioni approfondite al *dataset* di riferimento. La parte teorica di questo capitolo, così come gli approfondimenti in Appendice B, sono tratti da Newman (2018) e Avrachenkov and Dreveton (2022).

2.1 Concetti generali

2.1.1 Definizioni

Una rete viene definita come un insieme di nodi o vertici, uniti mediante degli archi. In base alla disciplina si potrebbero incontrare termini differenti, ad esempio in sociologia si parla di attori e legami.

Una rete può essere rappresentata tramite un grafo G composto dalla coppia (V, E), dove V è l'insieme finito di nodi ed E è l'insieme ordinato di coppie di nodi collegati tra loro $\{i, j\} \in E$, ossia l'insieme di archi. Graficamente i nodi sono rappresentati da punti, mentre le connessioni tra essi da frecce o linee.

La maggior parte delle reti è caratterizzata dalla presenza di un solo arco tra coppie di nodi e dall'assenza di nodi collegati con se stessi (*self-loops*). Si parla in questo caso di reti o grafi semplici. Qualora, invece, fossero presenti più archi per una qualsiasi coppia di nodi, la rete corrispondente è definita come multi-grafo (*multigraph*) e le connessioni tra due nodi vengono chiamate archi multipli (*multiedge*).

La rappresentazione matematica di una rete è data dalla matrice di adiacenza. Una matrice di adiacenza $\mathbf{A}$ è una matrice quadrata di dimensione $n \times n$, dove n è il numero di archi, i cui elementi A_{ij} sono pari a 1 se $\{i, j\} \in E$ e 0 altrimenti.

Se $A_{ij} = A_{ji}$ il grafo è indiretto, l'arco è rappresentato da una linea e la matrice di adiacenza è simmetrica. Di conseguenza, se esiste l'arco che va da i a j allora esiste lo stesso arco con direzione opposta. Se questa condizione non è verificata, si parla di grafo diretto (*digraph*), in cui le connessioni vengono rappresentate da frecce e la risultante matrice di adiacenza risulta asimmetrica. In questo caso, l'esistenza di un arco da i a j non implica che ne sia uno da j ad i.

Molte reti presentano archi che indicano la presenza o l'assenza di una connessione. In questi casi, $A_{ij} \in \{0, 1\}$ e la rete viene chiamata binaria. In altri situazioni può essere disponibile un'informazione sulla

forza del legame tra coppie di nodi, solitamente rappresentata da un numero reale positivo. In tal caso si utilizza il termine reti pesate e gli elementi della corrispondente matrice di adiacenza, A_{ij}, corrispondono al peso della connessione tra i e j. La diagonale della matrice di adiacenza sarà nulla in assenza di *self-loops*.

2.1.2 Statistiche descrittive

Per reti di dimensioni ridotte, una semplice visualizzazione grafica potrebbe, sebbene non sempre, rivelarsi sufficiente per estrarre informazioni utili. Tuttavia, con l'aumentare del numero di nodi, è preferibile utilizzare misure che descrivono le caratteristiche rilevanti della rete.

Tra gli indicatori più utilizzati vi è il grado. In una rete indiretta, il grado di un nodo i, k_i, è il numero di connessioni che esso presenta. Utilizzando i termini della matrice di adiacenza, k_i, in una rete con n nodi, può essere espresso come

$$k_i = \sum_{j=1}^{n} A_{ij}.$$

Nel caso di una rete diretta, ogni nodo può avere due tipologie di gradi. Il grado in entrata del nodo i, k_i^{in}, è il numero di connessioni verso i, mentre il grado in uscita del nodo i, k_i^{out}, è il numero di archi che partono da i. Quindi, considerando che $A_{ij} = 1$ se è presente un arco che parte da i e arriva a j, il grado in entrata e quello in uscita risultano rispettivamente

$$k_i^{\text{in}} = \sum_{i=1}^{n} A_{ij}, \quad k_i^{\text{out}} = \sum_{j=1}^{n} A_{ij}.$$

Viene chiamata densità di una rete, indicata con ρ, la frazione di archi presenti sul totale degli archi possibili

$$\rho = \frac{m}{\binom{n}{2}} = \frac{2m}{n(n-1)}.$$

La densità assume valori tra 0 e 1 ($\rho \in [0, 1]$) e può essere vista come la probabilità che due nodi, presi in modo casuale, siano connessi. Se la densità, all'aumentare del numero di nodi, rimane diversa da 0, allora la rete viene detta densa. Al contrario, viene denominata sparsa una rete per cui, al crescere di n, sia la densità che la proporzione di elementi non nulli nella matrice di adiacenza tendono ad annullarsi. Nella maggior parte delle applicazioni si ha a che fare con reti sparse.

Viene definito *shortest path* o percorso geodesico, il cammino più corto che congiunge una coppia di nodi. La *shortest distance* o distanza geodesica tra due nodi i e j è, quindi, la lunghezza dello *shortest path*, ovvero il più piccolo valore r tale che $[\mathbf{A}^r]_{ij} > 0$, dove l'elevamento a potenza della matrice di adiacenza permette di identificare un cammino di lunghezza r che congiunge i e j. Potrebbe non esserci un percorso geodesico tra una coppia di nodi, qualora non fosse presente alcun arco che li connetta. Un esempio di questo tipo si riscontra nelle reti con più componenti, costituite da sotto-grafi. Tutti i nodi appartenenti a uno stesso sotto-grafo sono connessi tra loro, non esistono invece archi che colleghino nodi di componenti differenti. Per convenzione si dice che la distanza tra questi nodi è infinita.

Il concetto di percorso geodesico può essere utilizzato per definire diversi indicatori. Il primo di essi è il diametro, calcolato come la lunghezza del più lungo *shortest path*.

È poi possibile introdurre altre misure di centralità, che vanno oltre il semplice grado di un nodo. La *closeness centrality* misura la vicinanza di un nodo rispetto agli altri presenti in una rete.

Indicando con d_{ij} la *shortest distance* tra i nodi i e j, allora

$$\ell_i = \frac{1}{n-1} \sum_{j(\neq i)} d_{ij} \tag{2.1}$$

rappresenta la più breve distanza media tra il nodo i e ogni altro nodo della rete. In 2.1, risulta ragionevole l'esclusione della distanza di un nodo con sé stesso (d_{ii}), che è nulla per definizione e non modifica,

dunque, il calcolo della somma. Questo porta all'utilizzo del fattore $\frac{1}{n-1}$ anziché $\frac{1}{n}$.

Un valore ridotto di 2.1 viene attribuito ai nodi che, in media sono separati dagli altri da una distanza limitata. Quindi, più un nodo è centrale, più la quantità 2.1 sarà piccola. Questo porta a definire la *closeness centrality*, C_i, come

$$C_i = \frac{1}{\ell_i} = \frac{n-1}{\sum_{j(\neq i)} d_{ij}}.$$

Un'ulteriore misura di importanza di un nodo è data dalla *betweenness centrality*, che indica quanto un nodo si trovi di passaggio nei cammini che congiungono altre coppie. Valori elevati di questo indicatore portano ad identificare nodi con un'alta influenza all'interno della rete, grazie ad un maggior accesso alle informazioni che vi transitano. Inoltre, la rimozione di questi elementi porterebbe a interrompere la comunicazione tra molti altri nodi.

Indicando ora con n_i^{jk} il numero di percorsi geodesici da j a k che passanno tramite i e con g_{jk} il numero totale di *shortest path* da j a k, la *betweenness centrality* è definita come

$$x_i = \sum_{j,k} \frac{n_i^{jk}}{g_{jk}}, \tag{2.2}$$

dove, per convenzione il termine della somma è nullo se sia n_i^{jk} che g_{jk} sono pari a zero.

La *betweenness centrality*, a differenza degli altri indicatori di centralità, indica quanto un nodo sia di passaggio, più che quanto sia ben collegato. Dunque, potrebbero esserci nodi con valori elevati di *betweenness centrality* ma con un grado ridotto. Nodi con queste caratteristiche vengono chiamati *broker*.

2.1.3 Proprietà

Molte reti reali condividono alcune proprietà comuni, la prima delle quali viene chiamata effetto del mondo piccolo (*small world*). Questo fenomeno, derivante da un esperimento sul passaggio di lettere condotto da Milgram nel 1967, ha portato al concetto di sei gradi di separazione, dimostrando che il percorso che collega una coppia di nodi è più breve di quanto si pensi.

Prendendo in considerazione una rete indiretta e indicando con d_{ij} il cammino più breve che congiunge i nodi i e j, la distanza media tra i e ogni altro nodo sarà

$$\ell_i = \frac{1}{n} \sum_j d_{ij}.$$

A partire da questa quantità è allora possibile definire la media delle distanze tra nodi nella rete nel suo complesso come la media dei valori $\ell_i, \forall i$, vale a dire

$$\ell = \frac{1}{n} \sum_i \ell_i = \frac{1}{n^2} \sum_j d_{ij}. \tag{2.3}$$

L'effetto del mondo piccolo ipotizza quindi che la quantità ℓ sia ridotta.

Un'altra delle proprietà più importanti delle reti riguarda la distribuzione del grado. Richiamando quanto definito nella sottosezione 2.1.2, il grado di un nodo è il numero di connessioni che esso presenta. La distribuzione del grado di una rete, p_k, è la frazione di nodi che presenta un grado pari a k e indica, quindi, la frequenza con cui i nodi con differenti valori di grado si presentano nella rete. Può anche essere vista come la probabilità che, prendendo casualmente un nodo della rete, questo abbia un grado pari a k. La distribuzione del grado in molte reti reali presenta una forma monotona decrescente con una lunga coda a destra, indicando che la maggior parte dei nodi ha poche connessioni, mentre pochi nodi, chiamati *hub*, hanno molte connessioni. Tali distribuzioni vengono chiamate leggi di potenza. È però importante evidenziare che la distribuzione del grado non fornisce una

visione completa della struttura della rete. Infatti, possono esistere reti diverse tra loro ma con gli stessi valori del grado.

Non meno importante è la transitività di una rete. Una relazione tra due nodi viene intesa come connessione se esiste un arco che li congiunge. Questa relazione è transitiva se il legame tra i nodi i e j e tra j e k implica anche che i e m siano connessi. Questo concetto è spesso riassunto dalla frase "l'amico del mio amico è anche mio amico". Si avrà perfetta transitività qualora tutti i nodi all'interno delle componenti di una rete risultino connessi tra loro. In tal caso, le componenti della rete vengono chiamate *clique*. Questo non si verifica nella maggior parte delle reti ma, anche se la connessione tra i e j e tra j e k non ne comporta necessariamente una tra i e k, aumenta la probabilità che questo avvenga.

La transitività viene misurata mediante il coefficiente di *clustering* C definito come la proporzione di percorsi di lunghezza due che formano un triangolo rispetto al totale dei percorsi di lunghezza due. Si avrà $C = 1$ nel caso di transitività perfetta e $C = 0$ in assenza di triadi chiuse.

Un'altra caratteristica che viene riscontrata nella maggioranza delle reti è la tendenza dei nodi a connettersi ad altri simili in termini di caratteristiche. Questa propensione viene chiamata omofilia o mescolamento assortativo (*assortative mixing*). È molto raro riscontrare il contrario, detto mescolamento disassortativo (*disassortative mixing*), ovvero il caso in cui nodi tendano a connettersi con altri da loro differenti.

2.2 Tipologie di reti

Esistono diverse tipologie di reti che possono essere utilizzate per rappresentare relazioni tra entità in contesti differenti. Le reti dirette, indirette, pesate e i multi-grafi citati in precedenza sono solamente alcune delle strutture disponibili. In questo paragrafo verrà posta

l'attenzione sulle reti bipartite, caratterizzate da due tipologie distinte di nodi, e sulle reti dinamiche, utilizzate per lo studio dell'evoluzione delle relazioni nel tempo.

2.2.1 Reti bipartite

I nodi possono essere connessi sulla base della loro appartenenza a gruppi o dalla partecipazione a specifici eventi, dando origine a quello che viene definito un legame di affiliazione. In sociologia si fa riferimento a questa struttura con il termine rete di affiliazione.

I grafi bipartiti, o bimodali, rappresentano una categoria generale, di cui le reti di affiliazione sono un caso specifico. Una rete bipartita è caratterizzata dalla presenza di due tipologie distinte di nodi e le connessioni possono avvenire solo tra nodi di natura differente. Di conseguenza, le strutture di rete descritte in precedenza in questo capitolo, composte da un solo tipo di nodo, vengono classificate come reti unipartite.

Un grafo bipartito viene definito come $G_B = (V_1, V_2, E)$ dove V_1 e V_2 sono i due diversi insiemi di nodi ed E è l'insieme degli archi. In generale, una rete bipartita è indiretta e, dunque, simmetrica. Ad esempio, considerando la partecipazione ad eventi da parte di un gruppo di persone, si può affermare che se una persona ha partecipato a un evento, quell'evento ha visto la presenza di quella persona.

La rappresentazione matematica di una rete bipartita (indiretta e non pesata) è la matrice di incidenza. Supponendo che V_1 e V_2 abbiano rispettivamente cardinalità n e g, il generico elemento della matrice di incidenza, che avrà dimensione $n \times g$, risulta

$$B_{ij} = \begin{cases} 1 \text{ se } i \text{ è affiliato a } j, \\ 0 \text{ altrimenti.} \end{cases}$$

È sempre possibile trasformare un grafo bipartito in uno unipar-

tito, collegando direttamente i nodi che appartengono al medesimo gruppo. Questo processo prende il nome di proiezione monomodale. Ogni rete bipartita avrà due proiezioni, una per ogni tipologia di nodi. Procedendo in questo modo, ogni gruppo della rete bipartita formerà un *clique* nella rete proiettata, un *cluster* in cui tutti i nodi sono connessi tra loro. Le rete proiettata risulta allora l'insieme di un numero di *clique* pari ai gruppi presenti nella rete bipartita.

Il passaggio dalla rete bipartita a quella proiettata comporta inevitabilmente una perdita di informazione. In particolare, non è possibile determinare il numero di gruppi condivisi tra le coppie di nodi. Una possibile soluzione consiste nell'effettuare una proiezione ponderata, in cui viene assegnato a ciascun arco che collega due nodi un peso corrispondente al numero di gruppi comuni a cui appartengono. Anche in questo modo non vengono preservate tutte le informazioni contenute nella rete bipartita di partenza, ma costituisce una rappresentazione più dettagliata rispetto a una proiezione semplice.

Matematicamente una rete proiettata P si ottiene moltiplicando la matrice di incidenza per la sua trasposta, $P = BB^T$, e avrà dimensione $n \times n$. L'elemento generico di P, che indica il numero di gruppi comuni tra i e j e quindi il peso attribuito alla connessione tra questi due nodi, sarà

$$P_{ij} = \sum_{k=1}^{g} B_{ik}B_{jk} = \sum_{k=1}^{g} B_{ik}B_{kj}^T$$

dove $B_{ik}B_{jk}$ sarà pari a 1 se i e j appartengono allo stesso gruppo.

La matrice P appare simile a una matrice di adiacenza di una rete unipartita pesata. L'unica differenza risiede nei valori P_{ii} che indicano il numero di gruppi a cui appartiene il nodo i e sono dunque diversi da zero anche in assenza di *self-loops*. Pertanto, per ottenere la matrice di adiacenza della rete proiettata pesata, una volta calcolata $P = BB^T$, gli elementi sulla diagonale verranno sostituiti con zero. Per una versione non ponderata, sarà sufficiente impostare a 1 gli elementi diversi da zero al di fuori della diagonale principale.

In modo del tutto simile è possibile ottenere la rete proiettata dei gruppi o degli eventi tramite $P^T = B^T B$. Per P e P^T si utilizzano rispettivamente i termini matrice persona-a-persona e gruppo-a-gruppo (Breiger, 1974) o matrice di co-affiliazione e matrice di sovrapposizione degli eventi (Faust, 1997).

2.2.2 Reti dinamiche

Fino ad ora, non si è presa in considerazione la possibilità che le connessioni tra i nodi possano evolvere nel tempo, come invece avviene in molti contesti. Nonostante l'aspetto dinamico possa risultare importante, molte ricerche condotte si focalizzano su reti di tipo statico. Questo è dovuto a vari fattori, tra i quali la minore maturità delle metodologie per l'analisi delle reti dinamiche e l'elevato carico computazionale richiesto.

Per arrivare alla definizione di reti dinamiche, è necessario fare una distinzione tra reti certe, in cui i nodi e gli archi una volta creati rimangono stabili, e reti incerte, dove la presenza degli archi può variare nel tempo. Quest'ultima tipologia si suddivide a sua volta in reti statiche, la cui struttura non cambia nel tempo, e reti dinamiche, dove nodi e archi cambiano, modificandosi nel tempo. A seconda della natura delle variazioni che subiscono, le reti dinamiche possono essere:

- incrementali, caratterizzate dall'apparizione di nuovi nodi o archi;
- decrementali, in cui alcuni nodi o archi esistenti scompaiono;
- miste, che presentano una combinazione di entrambe le situazioni precedenti, con nodi e archi che possono sia apparire che scomparire nel tempo.

Una rete dinamica o temporale non è altro che un grafo indicizzato dal tempo t

$$G(t) = (V(t), E(t)),$$

dove $V(t)$ è l'insieme di nodi al tempo t ed $E(t)$ l'insieme degli archi tra i nodi inclusi in $V(t)$. L'istante temporale t può variare in modo discreto o continuo all'interno di un determinato intervallo.

Il modo in cui viene misurata l'evoluzione di una rete dinamica può variare. Si potrebbe osservare precisamente l'apparizione o l'esclusione di archi e nodi. Al contrario, si potrebbe visualizzare solo un grafo statico G, che rappresenta un riassunto marginale dell'intera rete. Ci sono anche situazioni intermedie, come la conoscenza delle connessioni tra nodi all'interno di una finestra temporale limitata o l'utilizzo di istantanee della rete in intervalli di tempo specifici (Kolaczyk, 2020). L'ultima delle opzioni proposte è quella più comunemente utilizzata.

2.3 Applicazione ai dati

In questa sezione, si procederà all'analisi di rete utilizzando il *dataset OpenPowerlifting*. In particolare, ogni competizione può essere vista come un'opportunità di connessione tra gli atleti, dove la partecipazione comune ad eventi specifici genera archi che collegano i nodi rappresentati dagli atleti stessi. Verranno visualizzate le reti ed analizzate le statistiche descrittive introdotte in 2.1.2. Da notare che, anche se possono partecipare allo stesso evento competitivo, il fatto che uomini e donne non competano assieme rende necessario considerare questi eventi come separati.

La natura delle relazioni tra i nodi consente di considerare le reti anche in forma bipartita, distinguendo tra due tipologie di nodi: gli atleti e le competizioni. Inoltre, è possibile sfruttare la struttura delle reti dinamiche per visualizzare le connessioni tra gli atleti in intervalli temporali specifici, analizzando l'evoluzione delle relazioni all'interno della rete anno dopo anno.

Una volta conclusa l'analisi esplorativa è dunque necessario procedere alla creazione delle matrici di adiacenza. Queste vedranno disposti sia in riga che in colonna gli atleti e il generico valore A_{ij}^t sarà

pari a 1 se l'atleta i ha partecipato almeno ad una competizione con l'atleta j nell'anno t. Viene di seguito riportato un esempio di matrice di adiacenza per la rete femminile per $t = 1979$, il primo anno a disposizione.

$$A_t = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}.$$

È stato preso in considerazione questo anno per l'esiguo numero di atlete. La matrice ha infatti dimensione 7×7 e consente dunque di essere visualizzata nella sua interezza. Si noti che i valori lungo la diagonale principale sono nulli e non sono pari al numero di competizioni a cui ha preso parte ogni atleta, come avviene tramite la proiezione monomodale della matrice di incidenza del grafo bipartito.

A seguito della costruzione delle matrici di adiacenza binarie per i diversi anni a disposizione è possibile visualizzarle mediante dei grafi. Si concentra innanzitutto l'attenzione verso le atlete di sesso femminile e per queste vengono mostrate le reti per i primi (Figura 2.1) e gli ultimi sei anni a disposizione (Figura 2.2). Risulta evidente l'iniziale numerosità ridotta, principalmente dovuta al fatto che è proprio attorno a questi primi anni che il *Powerlifting* iniziò a diffondersi nel mondo femminile.

Nel 1980, 1981 e 1983 si osservano delle reti sconnesse. In particolare, nel 1980 si possono identificare due componenti. Nella componente più grande, tutti i nodi risultano connessi tra loro, formando una rete coesa. Al contrario, l'altra componente è costituita da un singolo nodo isolato, che non presenta alcuna connessione con gli altri atleti.

Interessante è anche la rete nell'anno 1982, in cui è chiara la presenza di un'unica atleta che ha partecipato a tutte e tre le competizioni tenutasi quell'anno. Questo nodo non solo appare in molti percorsi tra coppie di altri nodi, ma funge anche da ponte tra due gruppi che altrimenti non sarebbero connessi. Il suo ruolo di centralità viene infatti sottolineato anche dal valore di *betweenness*, pari a 25. Analizzando più nello specifico questa atleta, si vede che ha effettuato un cambio di categoria tra le tre gare che ha svolto, passando dalla categoria 63kg alla 56kg, in occasione della *US Military World Championships*. Valutando anche il peso registrato in occasione della competizione, si nota che è sufficientemente vicino al limite superiore della categoria, il che ha portato l'atleta ad essere molto competitiva, tanto da guadagnarsi il primo posto.

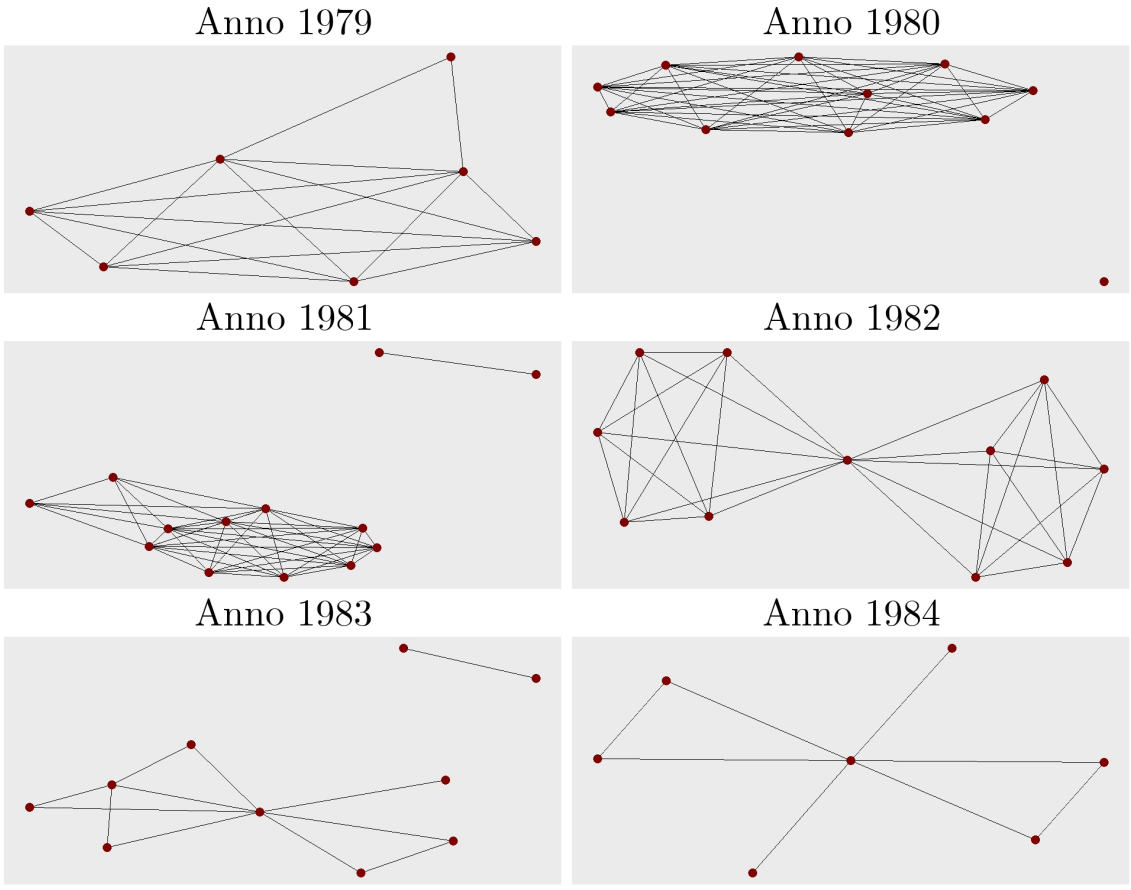


Figura 2.1: Reti di partecipazione competitiva femminile dal 1979 al 1984.

Esaminando gli ultimi sei anni a disposizione ci si aspetta, come viene confermato visivamente dai grafi in Figura 2.2, che il numero di nodi e di connessioni aumentino notevolmente. Le strutture sono molto più complesse rispetto ai primi anni a disposizione, con un numero decisamente più elevato di partecipazioni e connessioni. Anche in questo caso vengono riscontrate reti sconnesse, in particolare per gli anni 2021 e 2023 mentre nel 2024 si osservano due componenti. Quest'ultima situazione è dovuta al fatto che, al momento della raccolta dei dati, si erano svolte soltanto due competizioni. Un altro aspetto rilevante, che si noterà anche in seguito, è la diminuzione di connessioni registrata nel 2021, dovuta molto probabilmente alla pandemia di COVID-19.

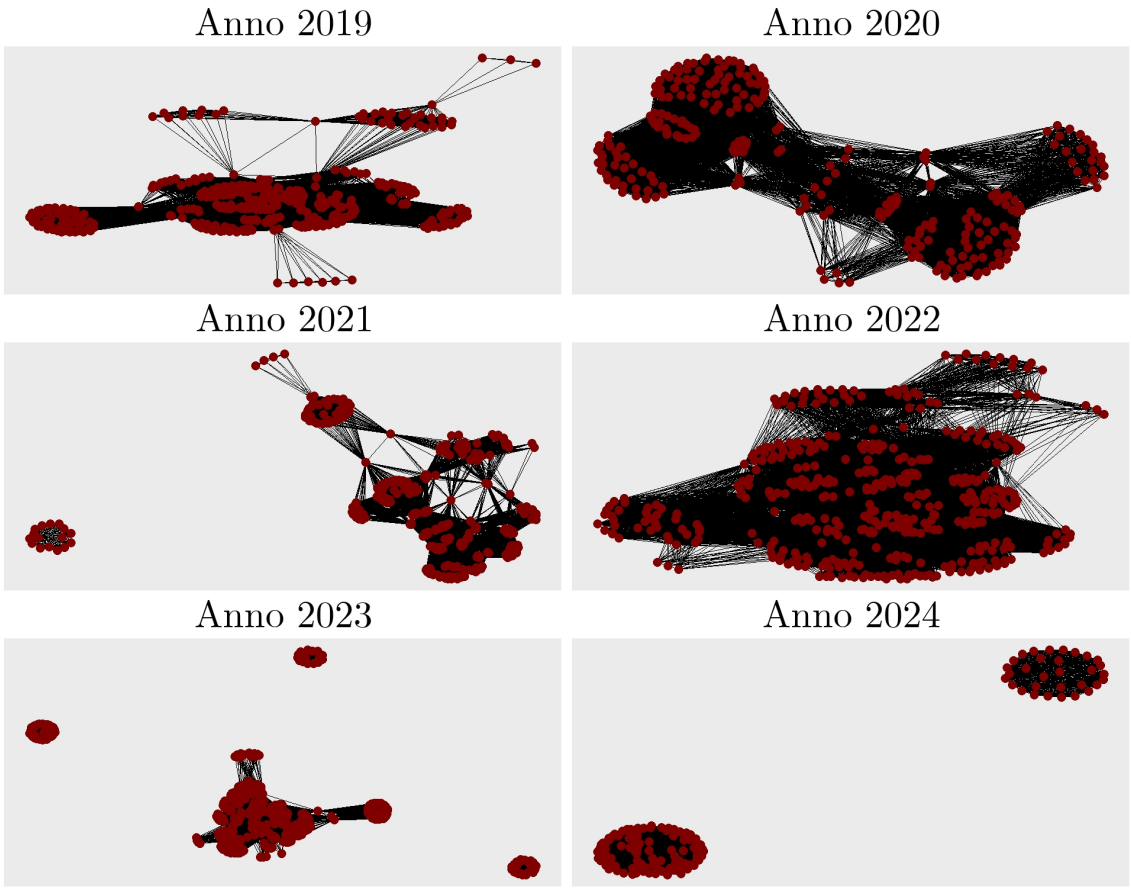


Figura 2.2: Reti di partecipazione competitiva femminile dal 2019 al 2024.

Si considerano ora alcune statistiche descrittive, quali la media del

grado, la densità e il diametro, e si visualizza come queste cambino nel tempo. Si nota innanzitutto che, tra il 1984 e il 2000, non tutti gli anni hanno visto lo svolgimento di competizioni. In particolare, nel periodo compreso tra il 1990 e il 2000, si è registrato un drastico calo nel numero di competizioni organizzate.

Dal grafico della densità (Figura 2.3) si evidenzia che le reti nel 1990, 1996 e 2000 hanno un valore di questa statistica pari a 1. Questo significa che il numero di archi osservati è pari a tutte le connessioni possibili ed è dovuto al fatto che è stata svolta una sola competizione (*WPC World Championships* nel 1990, *Women's European Powerlifting Championships* nel 1996 e *World Bench Press & Deadlift Championships* nel 2000). Eccetto per questi tre anni, la rete con maggiore densità è quella relativa al 1980, il cui grafo è stato riportato nella Figura 2.1. A seguito di un *trend* essenzialmente decrescente dal 2009 al 2023, si nota un picco verso l'alto nel 2024, nonostante, come evidenziato in precedenza, la numerosità delle competizioni per quest'anno risulta ridotta. Tuttavia, la presenza di due componenti all'interno delle quali ogni nodo è collegato a tutti gli altri ha determinato un notevole aumento della densità.

Il diametro, il cui andamento è riportato nella Figura 2.4, sottolinea la ridotta dimensione delle reti per gli anni a disposizione. Infatti, la lunghezza del cammino più lungo che congiunge due nodi è al massimo pari a 5, in corrispondenza del 2021.

Infine, la media del grado (Figura 2.5) è rimasta pressoché costante, con valori ridotti fino al 1984, si nota poi una crescita che porta ad un picco nel 1990, registrando una media superiore a 60. Successivamente una decrescita è seguita da un aumento fino al 2019. La media del grado è poi diminuita notevolmente passando da 110, nel 2019, a 88, nel 2020 e 61, nel 2021. Questa diminuzione, visualizzata anche nel grafo della rete di quell'anno, potrebbe essere attribuita al periodo di pandemia di COVID-19 che ha portato ad una diminuzione anche dei partecipanti alle gare.

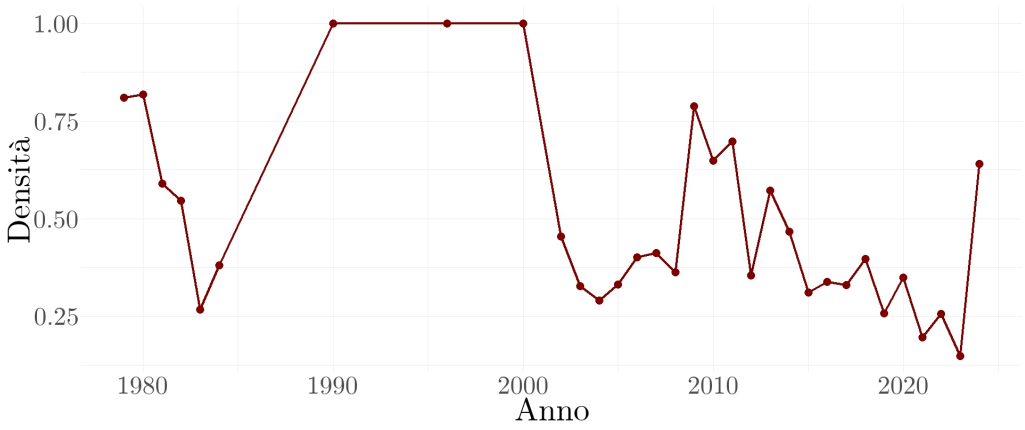


Figura 2.3: Densità delle reti femminili dal 1979 al 2024.

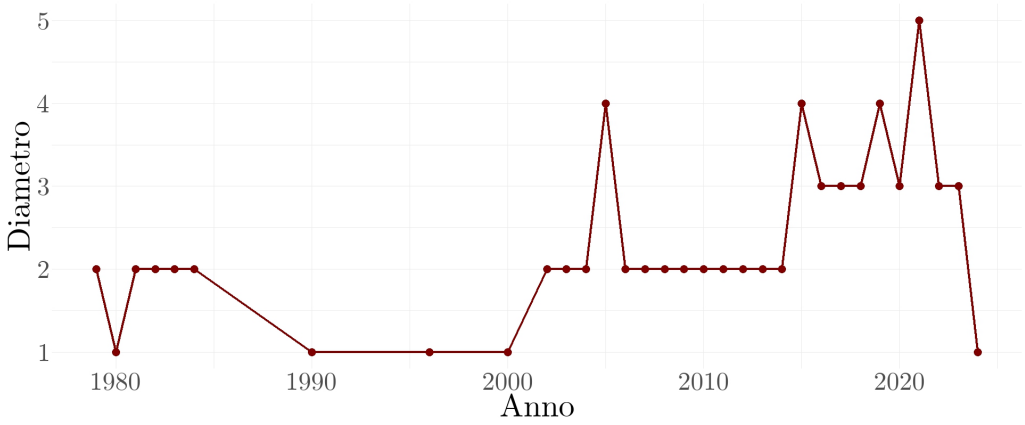


Figura 2.4: Diametro delle reti femminili dal 1979 al 2024.

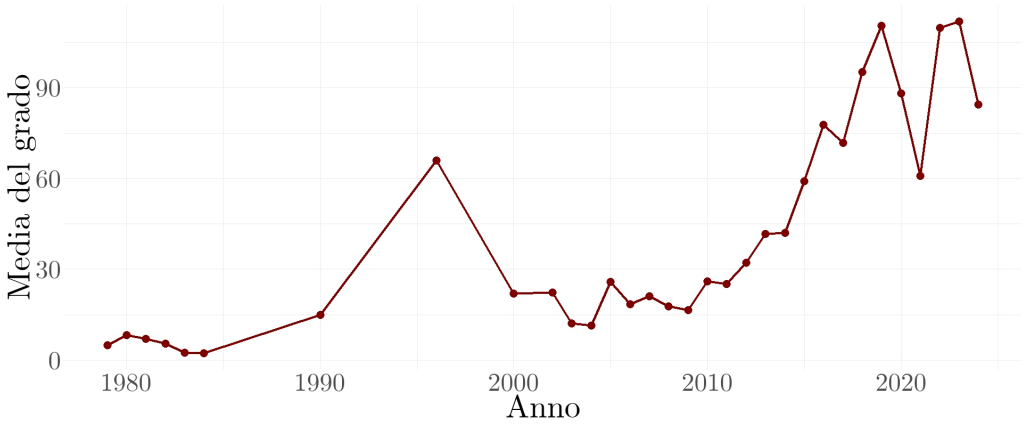


Figura 2.5: Media del grado delle reti femminili dal 1979 al 2024.

Concentrandosi sulla rete più recente con dati completi per l'intero anno, ossia quella del 2023, è possibile esaminare se la distribuzione del grado segue una legge di potenza.

Come avviene nella maggior parte delle reti, si osserva una maggiore frequenza di nodi con un numero relativamente ridotto di connessioni rispetto a nodi altamente connessi. Inoltre, la distribuzione del grado mostrata nella Figura 2.6 rappresenta una legge di potenza, anche se si nota un iniziale aumento della frequenza di nodi con valori ridotti del grado. Tuttavia, come specificato nella sottosezione B.1.1, anche distribuzioni che mostrano tali caratteristiche vengono considerate appartenenti a questa categoria.

Una volta analizzato il grado dei nodi nella rete, è possibile esaminare altri indicatori di centralità che offrono ulteriori spunti sul ruolo e sull'influenza di ciascun nodo.

La *closeness centrality* di questa rete risulta relativamente piccola, a sottolineare la presenza di percorsi più lunghi che portano i nodi ed essere distanti dagli altri appartenenti della rete. Questo si nota anche dal grafico del diametro (Figura 2.4), che mostra il suo valore più elevato proprio in corrispondenza del 2023.

L'istogramma della *betweenness centrality* in Figura 2.7 rivela valori sorprendentemente elevati di questo indicatore. Si specifica che il grafico non mostra 8 atleti per i quali questa statistica è particolarmente alta (superiore a 5000). Sebbene non siano numerosi, l'alta *betweenness centrality* permette di identificare atleti che fungono da ponte in molti percorsi tra coppie, indicando così anche una maggiore partecipazione alle competizioni. Si tratta quindi di soggetti con un'esperienza superiore e che potrebbero risultare molto competitivi. L'individuazione di questi nodi è importante non solo per comprendere la dinamica della rete, ma anche per gli altri atleti e i loro preparatori. Studiare le loro prestazioni può fornire informazioni preziose sulle strategie e le tecniche di gara. Comprendere come questi atleti si allenino e affrontino le competizioni potrebbe aiutare gli allenatori a strutturare programmazioni più efficaci e strategie di gara che aumentino le probabilità di successo quando si affrontano questi concorrenti.

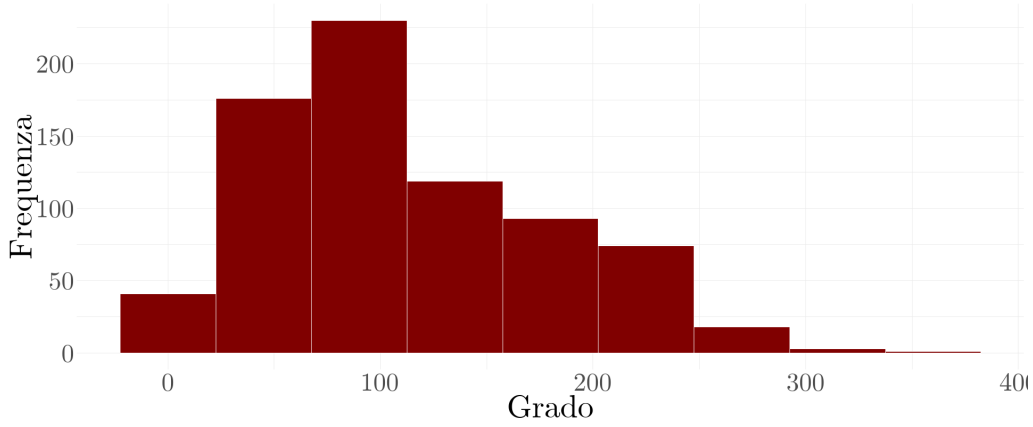


Figura 2.6: Distribuzione del grado della rete femminile del 2023.

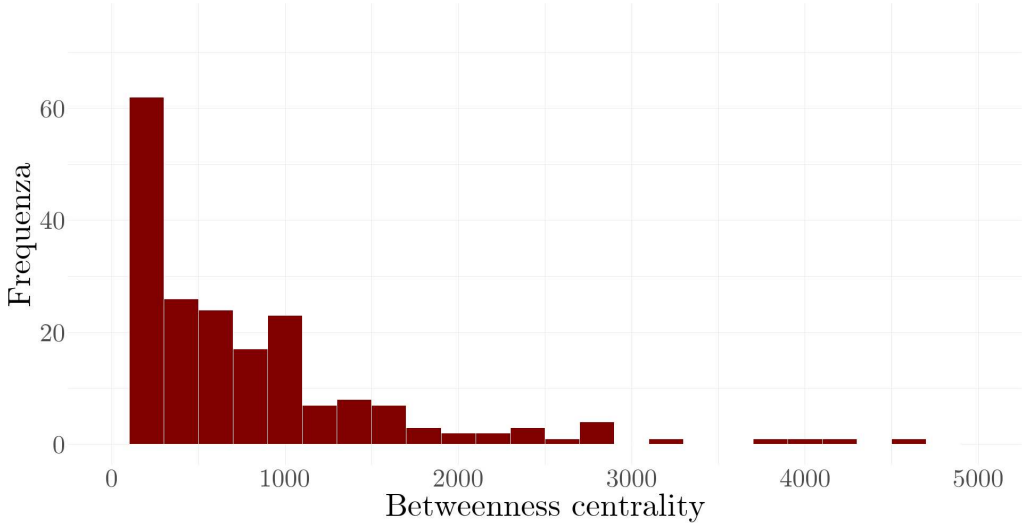


Figura 2.7: Distribuzione della *betweenness centrality* della rete delle partecipazioni competitive femminili del 2023.

Si passa ora all'analisi degli atleti di sesso maschile, per i quali vengono visualizzate le reti relative ai primi (Figura 2.8) e agli ultimi sei anni a disposizione (Figura 2.9), analogamente a quanto fatto per le atlete.

Una prima differenza rispetto a quanto emerso nel caso del *Powerlifting* femminile nei primi anni è la maggiore presenza di atleti. In quegli anni, infatti, il *Powerlifting* maschile era già molto più diffuso e, di conseguenza, anche le partecipazioni alle competizioni erano più numerose. Un elemento degno di nota poi emerge nel 1981, anno in cui si distingue chiaramente un gruppo isolato di nodi all'interno della

rete. Questi atleti, che hanno partecipato alla *Men's European Powerlifting Championships*, non hanno svolto alle altre gare quell'anno, che erano principalmente a livello italiano o gare che vedevano coinvolti l'Italia e gli Stati Uniti. Gli atleti che hanno partecipato ai campionati europei hanno quindi preso la decisione di non svolgere altre competizioni, per concentrarsi su quella che era considerata la competizione più importante dell'anno. Un altro aspetto evidente è l'incremento di *hub*, in particolare negli anni 1980, 1983 e 1984.

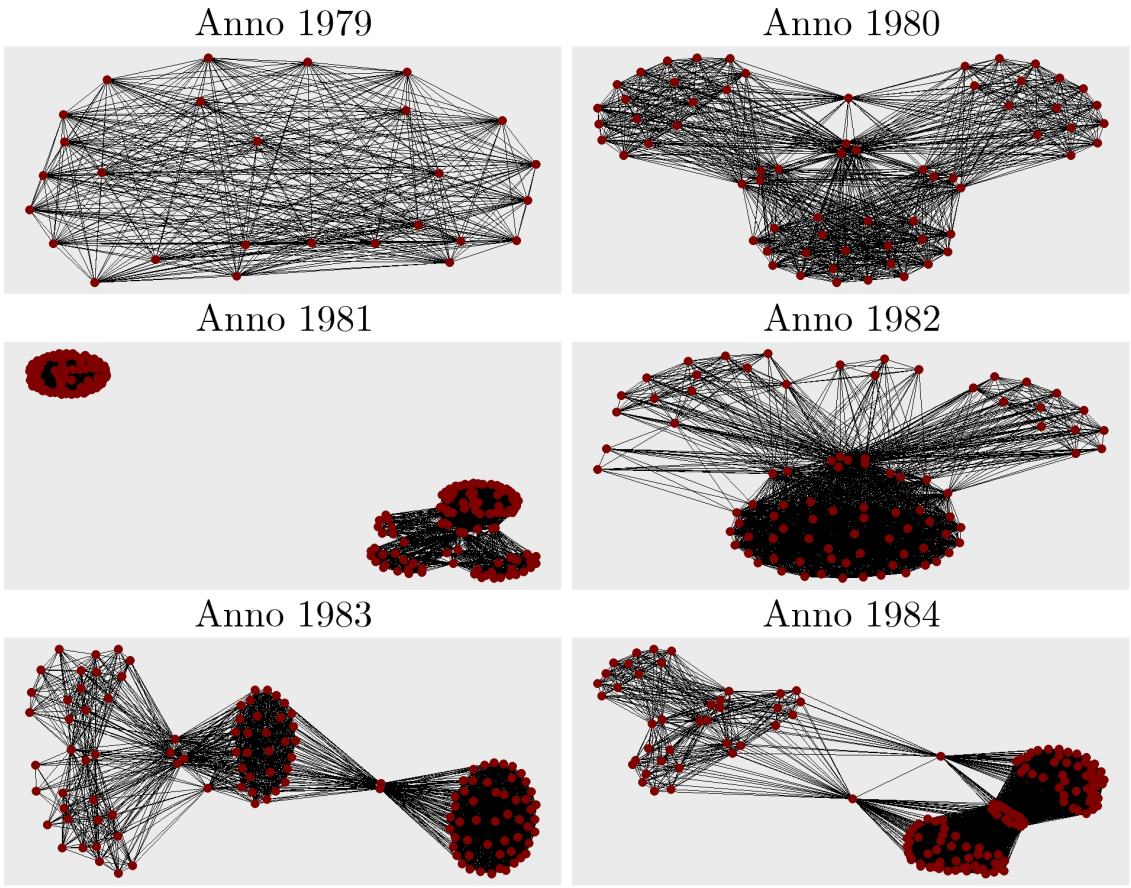


Figura 2.8: Reti di partecipazione competitiva maschile dal 1979 al 1984.

Similmente a quanto osservato per le competizioni femminili si può notare anche in questo caso una diminuzione della partecipazione nel 2021. Una differenza rispetto a quanto visto in precedenza per l'ambito femminile è invece quanto emerge nel 2024. Prima si erano distinti due gruppi a sé stanti e nessuna atleta, quindi, aveva partecipato ad entrambe le competizioni svolte ma solamente ad una delle due. In questo caso, invece, sono presenti due atleti che connettono i due gruppi. Le due competizioni svolte coinvolgono una sola alzata (distensioni su panca). È infatti meno probabile che, come si è verificato per il *Powerlifting* femminile, un atleta partecipi a gare che coinvolgono tutte e tre le alzate in tempi così ravvicinati.

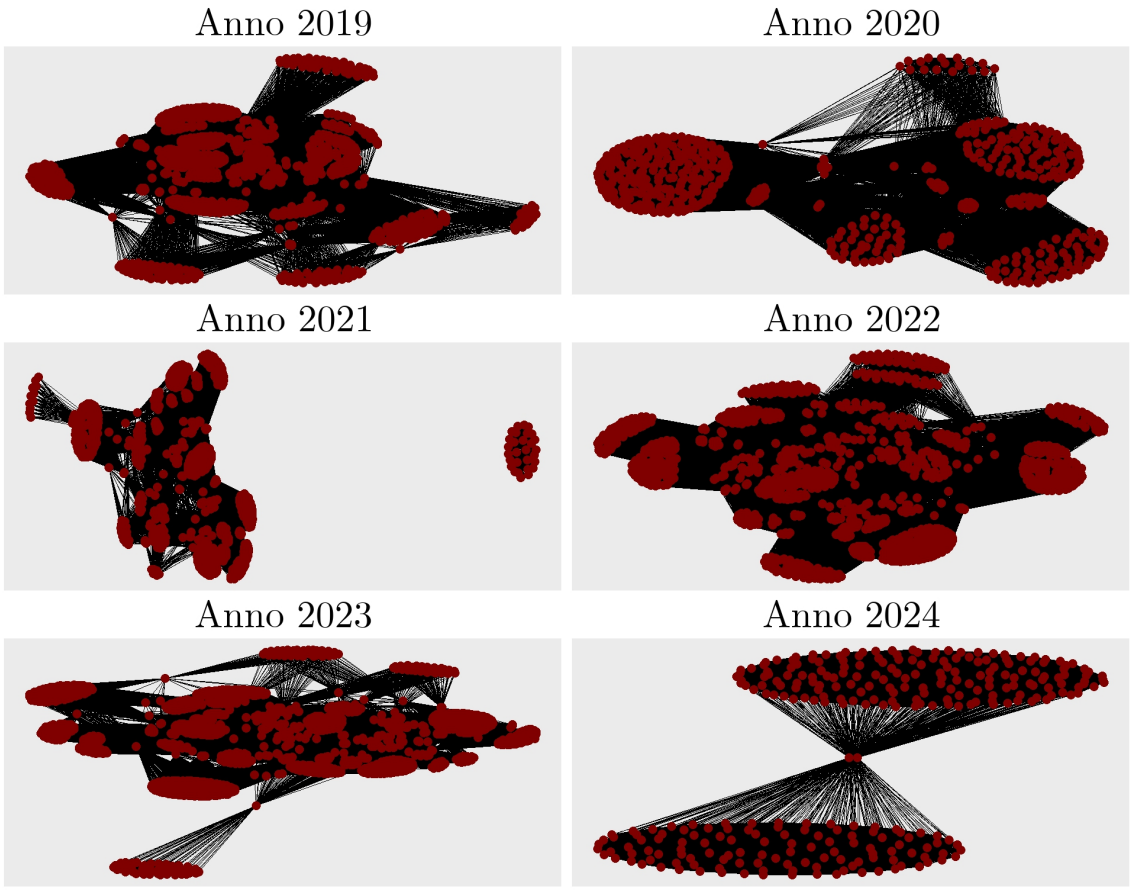


Figura 2.9: Reti di partecipazione competitiva maschile dal 2019 al 2024.

Si riportano le statistiche descrittive per il *Powerlifting* maschile, sovrapponendole ai grafici in Figura 2.3, 2.4 e 2.5 per rendere il confronto più diretto. Fino al 1982, la densità risulta inferiore per gli uomini rispetto alle donne. Questo indica che in quel periodo, nonostante la minor partecipazione di atlete nelle competizioni, le reti femminili erano più coese tra loro rispetto a quelle maschili. Il fatto che una rete abbia un maggior numero di nodi non implica infatti anche una maggiore densità, come si è verificato in questo caso. Dal 1985 al 2000 la densità delle reti maschili è pari a 1, indicando lo svolgimento di una singola competizione con conseguenti connessioni tra ogni coppia di nodi. A seguito degli anni 2000, gli andamenti appaiono più simili, mostrando in entrambi i casi una diminuzione a seguito della pandemia e una ripresa a partire dal 2023.

Analizzando il diametro si osserva una maggiore differenza tra i due andamenti, con picchi occasionali verso l'alto, in corrispondenza di competizioni di maggiore importanza, come il *World Championships* nel 1990 o le numerose gare a livello nazionale tenutasi nel 2021. In entrambi i casi, comunque, il diametro non assume valori elevati: il valore massimo osservato per le donne è 5 nel 2021, mentre per gli uomini è 4 nel 2005, 2015 e 2019.

Infine, la media del grado mostra un andamento simile per entrambi i sessi, con un picco nel 1990 e una diminuzione a seguito della pandemia. È evidente, inoltre, che i valori per gli uomini siano uniformemente più alti di quelli osservati per le donne. Questo è comprensibile, come già sottolineato più volte, in quanto, nonostante la crescente partecipazione in ambito femminile, il *Powerlifting* maschile ha sempre visto una maggiore partecipazione.

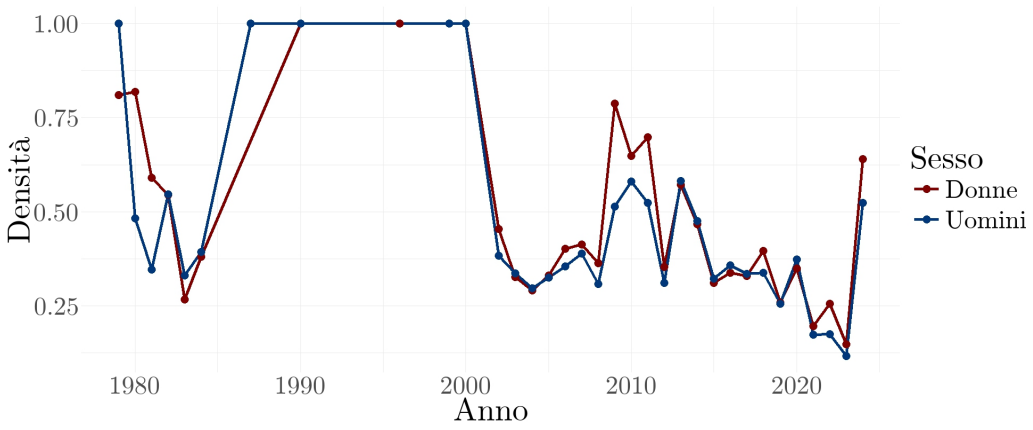


Figura 2.10: Densità delle reti femminili e maschili dal 1979 al 2024.

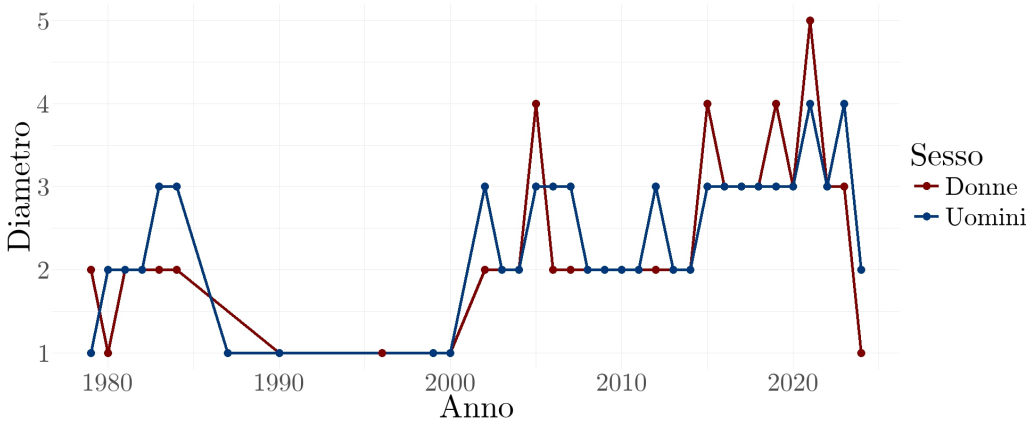


Figura 2.11: Diametro delle reti femminili e maschili dal 1979 al 2024.

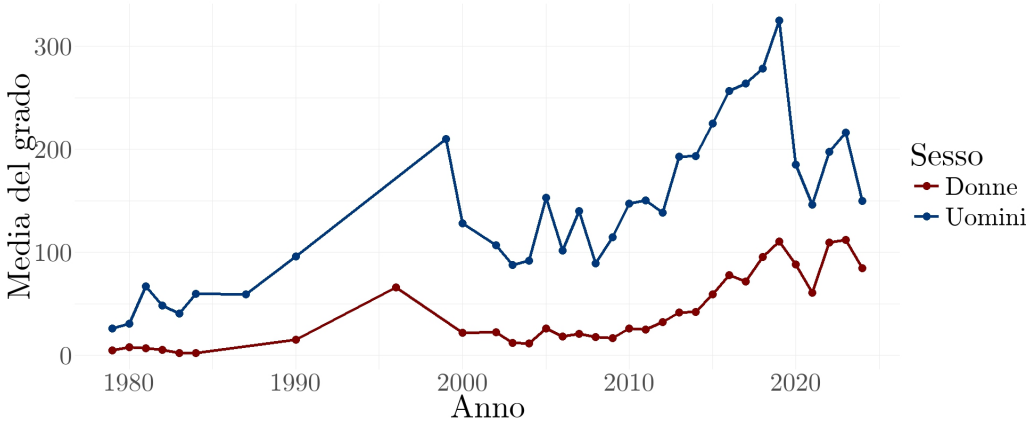


Figura 2.12: Media del grado delle reti femminili e maschili dal 1979 al 2024.

A differenza di quanto osservato per le atlete di sesso femminile, la distribuzione del grado nelle reti maschili del 2023 non segue una tipica legge di potenza. In particolare, per i valori ridotti del grado (fino a 200), si nota un aumento della frequenza, seguito da una diminuzione che, tuttavia, porta un secondo picco attorno al valore di 300. A partire da questo punto, la distribuzione riprende la sua natura monotona decrescente. È chiaro che la magnitudine differisca notevolmente dal grafico in Figura 2.6 relativo alle atlete, dove valori di grado così elevati erano presenti solo nella coda della distribuzione. La presenza di un numero maggiore di nodi con un grado elevato nelle reti maschili riflette la partecipazione più ampia, che porta a un maggior numero di connessioni e, di conseguenza, a valori di grado più elevati rispetto alle reti femminili.

Una situazione analoga a quella osservata per il *Powerlifting* femminile si riscontra nei valori di *closeness centrality*, che rimangono pressoché nulli, mentre i valori di *betweenness centrality*, come evidenziato dalla distribuzione mostrata nella Figura 2.14, continuano a essere molto elevati, con ben 47 atleti (esclusi dal grafico) che presentano un valore superiore a 10000. Un'ulteriore similitudine risiede nell'assenza di valori nulli di quest'ultimo indicatore, il che implica che ogni atleta è coinvolto in almeno un percorso tra coppie di altri nodi.

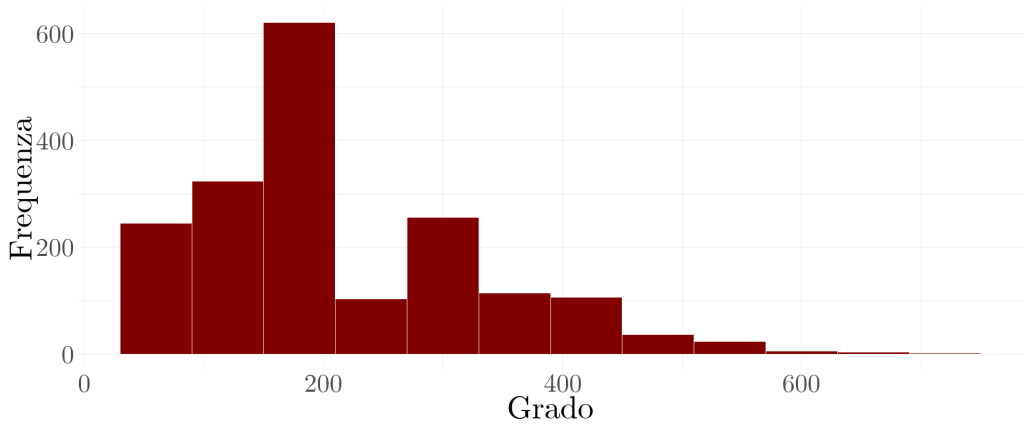


Figura 2.13: Distribuzione del grado della rete delle partecipazioni competitive maschili del 2023.

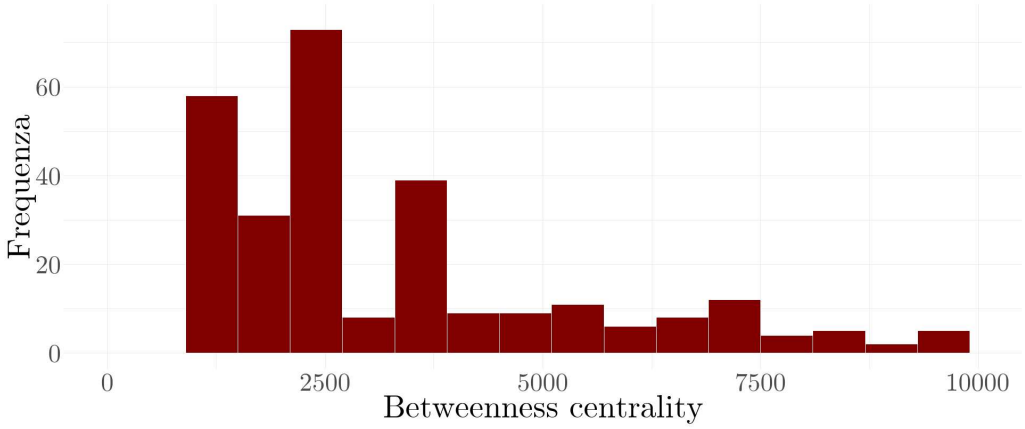


Figura 2.14: Distribuzione della *betweenness centrality* della rete delle partecipazioni competitive maschili del 2023.

Nella sottosezione 2.1.3 è stata descritta la proprietà di transitività di una rete, che comporta una tendenza a formare gruppi. Ci si aspetta un valore elevato del coefficiente di *clustering*, come avviene nella maggior parte delle reti. Infatti, l'utilizzo della proiezione monomodale della matrice di incidenza favorisce la formazione di gruppi all'interno della rete, rappresentata dagli eventi stessi. Nel 2023, il coefficiente di *clustering* per la rete di partecipazione competitiva maschile è risultato pari a 0.76, mentre per quella femminile è leggermente inferiore, con un valore di 0.7.

Dopo aver analizzato le reti, i loro cambiamenti nel tempo, così come le statistiche descrittive, si propone una breve panoramica sulle reti bipartite, al fine di esemplificare quanto illustrato in 2.2.1. Le matrici di adiacenza, impiegate per rappresentare le reti di partecipazione alle competizioni, sono in realtà proiezioni monomodali della matrice di incidenza del grafo bipartito.

La matrice di incidenza per le atlete del 1979 risulta essere la seguente:

$$B^T = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}^T .$$

La matrice di co-affiliazione $P = BB^T$ è quindi

$$P = \begin{bmatrix} 2 & 1 & 2 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 2 & 1 & 2 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 \end{bmatrix},$$

dove sulla diagonale è indicato il numero di competizioni a cui ogni atleta ha preso parte. Per ritornare alla formulazione della matrice di adiacenza A, riportata all'inizio di questo paragrafo, questi valori sulla diagonale vengono sostituiti con zero, poiché non si considerano archi riflessivi, e i valori maggiori di 1 al di fuori della diagonale principale vengono ridotti a 1, in modo da rappresentare una connessione binaria tra gli atleti.

Il grafo bipartito, riportato nella Figura 2.15, rappresenta le due tipologie di nodi: a destra le competizioni, corrispondenti alle colonne della matrice di incidenza B, e a sinistra le atlete, indicate nelle righe della stessa matrice. La visualizzazione risulta chiara grazie alla ridotta partecipazione registrata nel 1979 per il *Powerlifting* femminile. Vengono facilmente individuate le uniche due atlete che hanno preso parte a entrambe le competizioni. Queste corrispondono alla prima e alla terza riga della matrice di co-affiliazione B, il cui valore lungo la diagonale equivale esattamente a 2. Il grafo evidenzia inoltre una maggiore partecipazione alla competizione *Aviano Bench Press Championships*.

Per concludere, nella Figura 2.16 viene riportato il grafo bipartito di una rete più recente, quella del 2023. La matrice di incidenza per quell'anno ha una dimensione 755×22, ad indicare lo svolgimento di 22 competizioni che hanno coinvolto 755 atlete. Questo rappresenta un notevole aumento rispetto alla rete del 1979. È chiaro che questo

incremento di nodi e di eventi renda la visualizzazione e l'interpretazione del grafo più complessa. Risulta tuttavia chiara una maggiore partecipazione a più di un evento da parte delle atlete, evidenziando un coinvolgimento più attivo rispetto al passato. Inoltre, alcune competizioni mostrano un numero di partecipazioni significativamente più elevato.

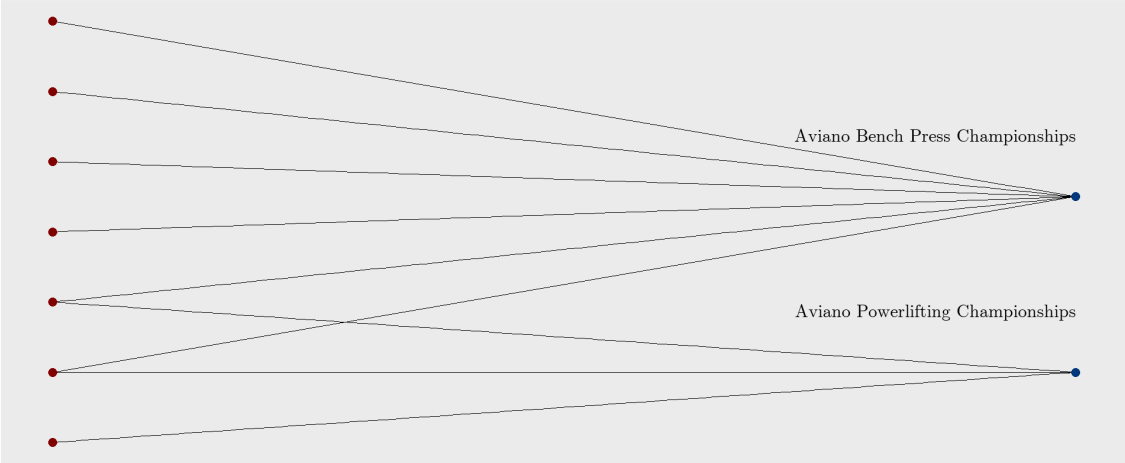


Figura 2.15: Grafo bipartito delle partecipazioni femminili nel 1979.

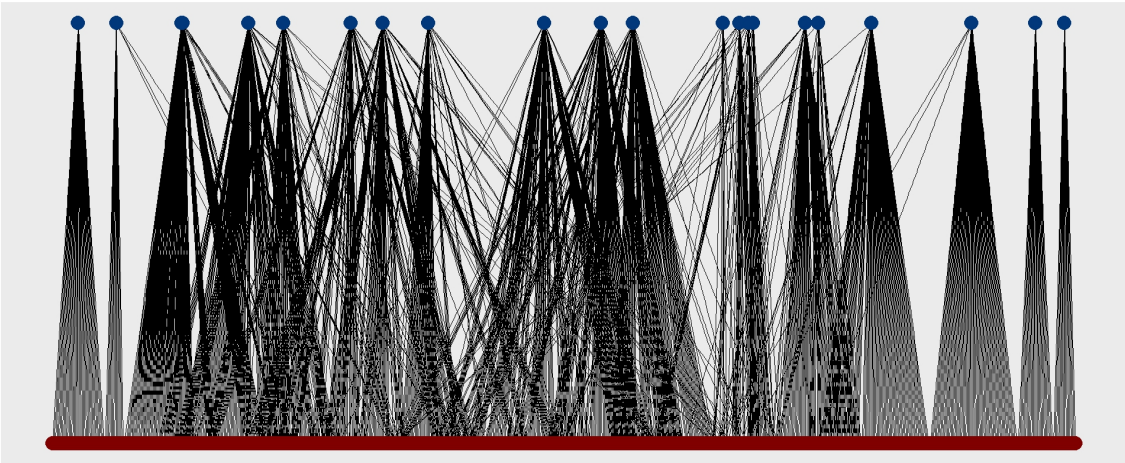


Figura 2.16: Grafo bipartito delle partecipazioni femminili nel 2023.

Capitolo 3

Visualizzazione delle reti

I grafi, come evidenziato nell'applicazione ai dati nel Capitolo 2, rappresentano uno strumento fondamentale per la visualizzazione e l'analisi delle reti. L'efficacia di questa rappresentazione è fortemente influenzata dalla sua estetica, che riguarda la disposizione dei nodi e, di conseguenza, degli archi. Un *layout* ben progettato facilita la comprensione delle dinamiche della rete e ne migliora l'interpretazione.

Il problema di ottimizzazione della visualizzazione delle reti è stato oggetto di numerosi studi, portando allo sviluppo di diversi metodi (Yifan, 2006). Tra questi, per reti semplici e non dirette, gli algoritmi *force-directed* sono tra i più flessibili e intuitivi (Kobourov, 2013). Un primo obiettivo in questo capitolo è quindi quello di illustrare i principali algoritmi che rientrano in questa classe.

Successivamente, si esploreranno le tecniche più utilizzate per il rilevamento delle comunità, che sono utili per identificare gruppi di nodi all'interno della rete. Analogamente a quanto avviene per i metodi di visualizzazione, anche il problema di *community detection* ha portato allo sviluppo di diversi algoritmi che ottimizzano metriche specifiche, tra cui la modularità, impiegata nell'algoritmo di Louvain, e la *betweenness*, sfruttata dall'algoritmo di Girvan-Newman.

Il capitolo si concluderà con l'applicazione e il confronto dei vari metodi discussi sulle reti di partecipazione competitive del *Powerlifting*.

3.1 Algoritmi *force-directed*

Si consideri una rete semplice indiretta con archi rappresentati da linee rette. Un algoritmo di *force-directed placements* affronta il problema del disegno di grafi mediante un modello fisico in cui i nodi sono rappresentati come corpi che interagiscono attraverso forze. I grafi generati da questi algoritmi tendono a mostrare una disposizione uniforme dei nodi, minimizzando gli incroci tra gli archi per favorire la simmetria e garantire una lunghezza omogenea degli archi stessi (Kobourov, 2013).

L'algoritmo di Fruchterman e Reingold (Fruchterman, 1991) si basa sul lavoro di Eades, la cui idea di base è quella di sostituire i nodi della rete con degli anelli in acciaio e gli archi tramite delle molle. Viene in questo modo formato un sistema meccanico governato da due forze. La forza repulsiva è presente tra ogni coppia di nodi ed è tanto più grande quanto più questi si trovano ad una distanza ridotta. La forza attrattiva, invece, coinvolge solamente coppie di nodi connessi, riducendo la complessità computazionale, ed è proporzionale al quadrato della distanza tra essi esistente. I nodi vengono rilasciati da una configurazione iniziale e saranno le due forze a portare alla minimizzazione dell'energia presente nel sistema. Il posizionamento finale dei nodi determina quindi un equilibrio tra la forza attrattiva e quella repulsiva, che si bilanciano fino ad annullarsi reciprocamente.

Un approccio alternativo è rappresentato dall'algoritmo Kamada-Kawai (Kamada, 1989), noto anche come modello a molla (*spring model*). Fondato sulla teoria dei grafi, questo algoritmo definisce la distanza ideale tra due nodi come quella teorica che li separa. Pertanto, la configurazione ottimale si raggiunge quando le distanze geometriche tra i nodi coincidono con quelle teoriche. Poiché frequentemente è difficile raggiungere tale obiettivo, l'algoritmo utilizza un sistema di molle per minimizzare l'energia, intesa come differenza tra le distanze geometriche osservate e quelle teoriche.

Ci sono due principali limitazioni associate all'uso degli algoritmi di *force-directed placements*. La prima è la presenza di molti minimi locali, che può portare l'algoritmo a fermarsi in configurazioni subottimali, soprattutto in grafi di grandi dimensioni. La seconda è la complessità computazionale. L'algoritmo di Fruchterman e Reingold calcola, ad ogni iterazione, le forze attrattive e repulsive, con un costo computazionale di $O(|E|)$ e $O(|V|^2)$ rispettivamente. Tuttavia, esiste una variante più efficiente che ignora la forza repulsiva tra nodi distanti. L'algoritmo di Kamada e Kawai, data la necessità di calcolare le distanze tra tutte le coppie di nodi, è computazionalmente più oneroso presentando un costo di $O(|V|^3)$ (Kobourov, 2013). Inoltre, le prestazioni di questi algoritmi tendono a peggiorare con l'aumentare delle dimensioni delle reti.

3.2 *Community detection*

L'identificazione delle comunità è uno degli obiettivi più comuni nello studio delle reti, ma può rivelarsi un compito complesso a causa della natura generale del problema. Lo scopo della *community detection* è quello di individuare suddivisioni della rete in insiemi di nodi che presentano molte connessioni al loro interno e poche verso l'esterno. Dunque, in questo contesto, una buona partizione viene intesa come una suddivisione in cui i nodi appartenenti a comunità distinte sono il più possibile separati, minimizzando le connessioni presenti tra esse. Tuttavia, è fondamentale chiarire ulteriormente cosa si intenda per "buona" partizione e come definire quantitativamente "molte" e "poche" connessioni (Newman, 2018).

Il metodo più diffuso consiste nell'utilizzare un indice che valuti le diverse suddivisioni generate, selezionando come migliore quella associata al punteggio più elevato. La metrica più comunemente utilizzata è la modularità, calcolata come la differenza tra la frazione di archi che collegano nodi appartenenti alla stessa comunità e il valore atteso

di questa quantità in una rete con connessioni casuali. Questa misura assume valori compresi tra 0 e 1 e raggiunge il valore massimo quando non sono presenti connessioni tra comunità diverse, indicando una netta separazione tra i gruppi di nodi. Tuttavia, è comune osservare valori che variano tra 0.3 e 0.7, mentre punteggi superiori sono rari (Newman and Girvan, 2004). Il calcolo della modularità per tutte le k^n possibili partizioni degli n nodi in k comunità è chiaramente non praticabile. Pertanto, la suddivisione trovata potrebbe non essere la migliore, ma può comunque fornire informazioni utili.

Tra gli algoritmi per la massimizzazione della modularità, il metodo di Louvain è senza alcun dubbio il più popolare. Questo algoritmo segue un approccio agglomerativo ed alterna due fasi. Nella prima fase, ogni nodo viene inizialmente trattato come una comunità a sé stante e viene spostato in quella che comporta il maggiore incremento della modularità. La seconda fase è analoga alla prima, ma in questo caso gli spostamenti riguardano i gruppi già formati. Ogni comunità del nuovo grafo viene dunque considerata come un singolo nodo. Il processo inizia nuovamente dalla prima fase e viene iterato fino a quando non si verificano più spostamenti e, di conseguenza, guadagni di modularità. L'algoritmo di Louvain si è dimostrato essere molto veloce ed accurato nel caso di reti statiche (Aynaud and Guillaume, 2010).

L'algoritmo di Girvan-Newman (Newman and Girvan, 2004), diversamente dal metodo di Louvain, segue un approccio divisivo. Invece di iniziare dai singoli nodi e aggregarli progressivamente in comunità, questo algoritmo le identifica rimuovendo iterativamente le connessioni esistenti nella rete. Un'ulteriore distinzione risiede nella metrica utilizzata, che in questo caso è la *edge betweenness*, definita come il numero di *shortest paths* che attraversano uno specifico arco. Le connessioni tra diverse comunità presenteranno un alto valore di *edge betweenness*. L'algoritmo inizia calcolando la *betweenness* per ogni arco della rete. Dopo aver rimosso quello con il valore più alto, relativo a con-

nessioni tra comunità, la *betweenness* viene ricalcolata per adattarsi alla nuova configurazione della rete. Il processo prosegue iterativamente, rimuovendo ogni volta l'arco con la *betweenness* più elevata. Il calcolo iterativo della *edge betweenness* comporta un aumento del costo computazionale, rendendo l'algoritmo meno efficiente rispetto a quelli basati sull'ottimizzazione della modularità, come l'algoritmo di Louvain (Newman, 2018).

Un altro metodo molto diffuso è l'algoritmo *InfoMap*. Questo algoritmo, basato sulla teoria dell'informazione, utilizza il concetto di *random walk*: una sequenza di nodi selezionati casualmente. Quando una rete presenta molti archi all'interno delle comunità rispetto a quelli che le collegano, ci si aspetta che una passeggiata casuale rimanga prevalentemente all'interno di ciascun gruppo. In altre parole, se ci sono poche connessioni tra le diverse comunità, una passeggiata casuale che attraversa questi archi richiederà un numero maggiore di passaggi, risultando così più lunga. Per quantificare questi percorsi, *InfoMap* associa a ciascuna passeggiata una stringa di *bit* composta da zero e uno. L'algoritmo identifica come partizione ottimale della rete quella associata alla stringa più corta, poiché una lunghezza ridotta indica che le passeggiate casuali tendono a rimanere all'interno delle comunità, suggerendo la presenza di gruppi più coesi.

Il principale vantaggio dell'algoritmo *InfoMap* è l'efficienza computazionale, con un tempo di esecuzione paragonabile a quello dell'algoritmo di Louvain. Risulta invece più efficiente dell'algoritmo Girvan-Newman. Inoltre, sebbene gli studi sull'efficacia delle tecniche di rilevamento delle comunità non abbiano individuato un algoritmo migliore in assoluto, i metodi basati sull'ottimizzazione della modularità e *InfoMap* sono spesso considerate tra le più performanti (Newman, 2018).

3.3 Applicazione ai dati

Nella sezione 2.3 è stato evidenziato come un elevato valore di *betweenness* possa rappresentare atleti più competitivi grazie ad una maggiore esperienza data da una partecipazione più continuativa alle competizioni. Una disposizione casuale dei nodi non permetterebbe di individuare efficacemente né questi atleti né comunità fortemente connesse presenti all'interno della rete. Gli algoritmi di *force-directed placements* precedentemente illustrati si rivelano quindi particolarmente utili per visualizzare le reti di partecipazione competitiva, poiché conservano informazioni derivanti dalla disposizione dei nodi, fondamentali per comprendere le dinamiche di interazione tra gli atleti e per identificare i ruoli strategici che alcuni di essi possono ricoprire all'interno della rete.

La presenza di comunità potrà essere rilevata tramite i metodi di *community detection*. Lo scopo principale sarà valutare l'esistenza di gruppi di atleti connessi che si distinguono per competitività in termini di prestazioni. Considerando la natura temporale delle reti, verrà adottato anche un approccio dinamico per il rilevamento delle comunità tramite l'algoritmo di Louvain, che permetterà di tenere conto della formazione e dissoluzione delle connessioni nel corso degli anni. La qualità delle configurazioni ottenute dalle reti statiche, rappresentate da istantanee di diversi anni, sarà confrontata con quella delle reti dinamiche. Il rilevamento delle comunità nella sua versione classica si concentrerà in particolar modo sulle reti di partecipazione competitiva femminili e maschili del 2023, l'anno più recente con dati completi sulle gare di *Powerlifting*.

3.3.1 Algoritmi *force-directed*

Prendendo in considerazione le reti femminili, a sinistra nella Figura 3.1 è rappresentata la rete con nodi disposti in modo casuale. Seb-

bene questa configurazione fornisca una prima visualizzazione del grafo, non facilita l'interpretazione delle dinamiche della rete e complica l'individuazione delle comunità. Al centro della figura, l'algoritmo Fruchterman-Reingold genera una configurazione più organizzata, evidenziando la presenza di tre componenti distinte. Tuttavia, si nota un'area molto coesa al centro della rete, che può rendere difficile l'analisi dei legami tra i nodi. Anche aumentando notevolmente il numero di iterazioni, passando da 500 a 5000, la configurazione rimane pressoché invariata. Ciò indica che l'equilibrio energetico della rete viene raggiunto già con un numero inferiore di iterazioni.

Infine, la disposizione dei nodi determinata dall'algoritmo Kamada-Kawai è mostrata nella parte destra della Figura 3.1. Rispetto al metodo Fruchterman-Reingold, si osserva una maggiore distanza tra i nodi. Anche in questo caso, si evidenziano alcune comunità di nodi meno numerose e più distanti dal gruppo centrale, le quali potrebbero riferirsi a competizioni di minore rilevanza. La sovrapposizione inferiore dei nodi nel terzo algoritmo consente di visualizzare più efficacemente le relazioni tra atleti e competizioni.

Focalizzandosi sulle reti maschili riportate in basso nella Figura 3.1 si nota una migliore disposizione dei nodi generata dall'algoritmo Fruchterman-Reingold, evidenziando una minore concentrazione rispetto al grafo femminile. La configurazione realizzata mediante l'algoritmo di Kamada e Kawai è caratterizzata, anche in questo caso, da una distanza maggiore tra i nodi, mostrando in modo più distinto la presenza di alcuni gruppi fortemente connessi.

Si noti inoltre come il grafo generato dall'algoritmo Fruchterman-Reingold consenta di identificare più di un nodo con ruolo di intermediario. Vengono quindi messi in luce elementi chiave che caratterizzano la rete, come la presenza di atleti centrali, di gruppi più distanti di atleti che hanno svolto una sola competizione oppure ancora di comunità più dense e fortemente connesse, che verrebbero totalmente persi mediante una semplice disposizione casuale dei nodi.

A causa della maggiore numerosità delle reti maschili, è opportuno considerare anche il tempo di esecuzione necessario per ottenere le configurazioni visualizzate. L'algoritmo Kamada-Kawai ha impiegato quasi cinque secondi, mentre Fruchterman-Reingold ha richiesto meno di mezzo secondo. Questi risultati indicano una maggiore efficienza del secondo metodo, come evidenziato anche nella sezione 3.1.

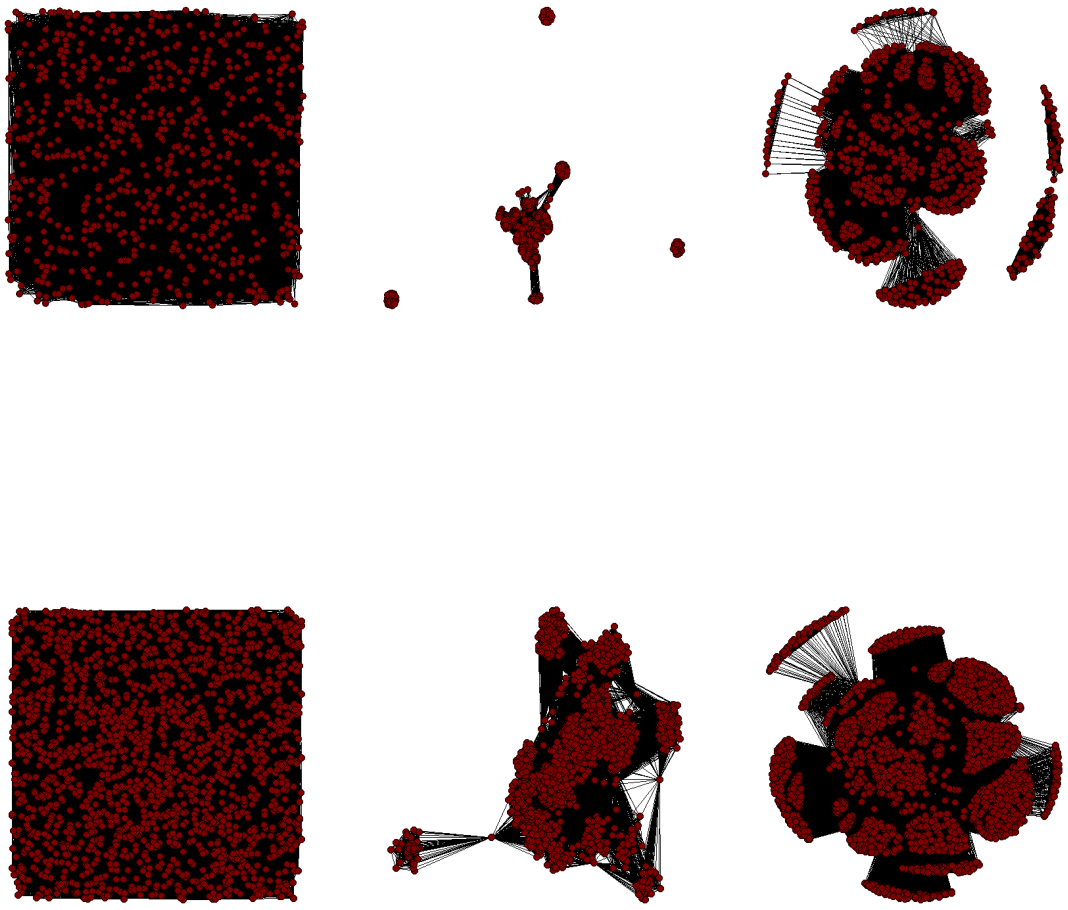


Figura 3.1: Reti femminili (in alto) e maschili (in basso) del 2023: disposizione casuale dei nodi (a sinitra), configurazione dell'algoritmo Fruchterman-Reingold (al centro) e dell'algoritmo Kamada-Kawai (a destra).

I grafi riportati in Figura 3.1 sono stati realizzati utilizzando `igraph` (Csárdi et al., 2024), una collezione di librerie per l'analisi delle reti disponibile in `R` e in altri *software*. Gli algoritmi utilizzati in questo elaborato rappresentano solo una parte delle soluzioni offerte dal pacchetto `igraph`. Per approfondimenti si rimanda alla sezione C.1.

3.3.2 *Community detection* statica

Si procede ora con il rilevamento delle comunità, analizzando innanzitutto le reti di partecipazione competitiva nel contesto del *Powerlifting* maschile. Le reti di ogni anno saranno inizialmente considerate come istantanee, escludendo la dinamica di formazione e scioglimento degli archi nel tempo.

Il grafico in Figura 3.2 mostra l'andamento della modularità delle comunità identificate dagli algoritmi di Louvain e *InfoMap*. È evidente la differenza tra le due tendenze in determinati periodi nei quali *InfoMap* mostra valori inferiori rispetto a Louvain. Le reti degli anni nei quali si notano le maggiori differenze risultano fortemente connesse, portando ad una maggiore difficoltà a rilevare comunità utilizzando *InfoMap*. Tuttavia, si osserva che nelle reti con i valori più elevati di modularità, gli andamenti tra i due algoritmi sono molto simili.

Inoltre, si evidenziano periodi in cui la modularità è nulla o quasi nulla per *InfoMap*, o presenta picchi negativi per l'algoritmo di Louvain. È importante ricordare che la modularità assume il suo valore massimo quando le comunità sono completamente separate, senza connessioni tra loro. Gli anni in cui si registrano partizioni peggiori in termini di modularità sono quelli caratterizzati dalla presenza di uno o più nodi intermediari, ossia con elevati valori di *betweenness*. Questi nodi, fungendo da ponti all'interno del grafo, pur essendo assegnati a una determinata comunità, mantengono numerose connessioni con gli altri gruppi, provocando la diminuzione di modularità osservata.

Infine, va sottolineata la rapidità di entrambi gli algoritmi, che hanno impiegato un massimo di quattro secondi per la rete del 2023. In termini di efficienza computazionale, l'algoritmo di Louvain si è dimostrato leggermente più veloce. Le grandi dimensioni delle reti di partecipazione competitiva nel *Powerlifting* maschile non hanno invece reso possibile l'utilizzo dell'algoritmo di Girvan e Newman, il cui costo computazionale si è rivelato eccessivo non consentendo di ottenere la suddivisione in comunità per i singoli anni a disposizione.

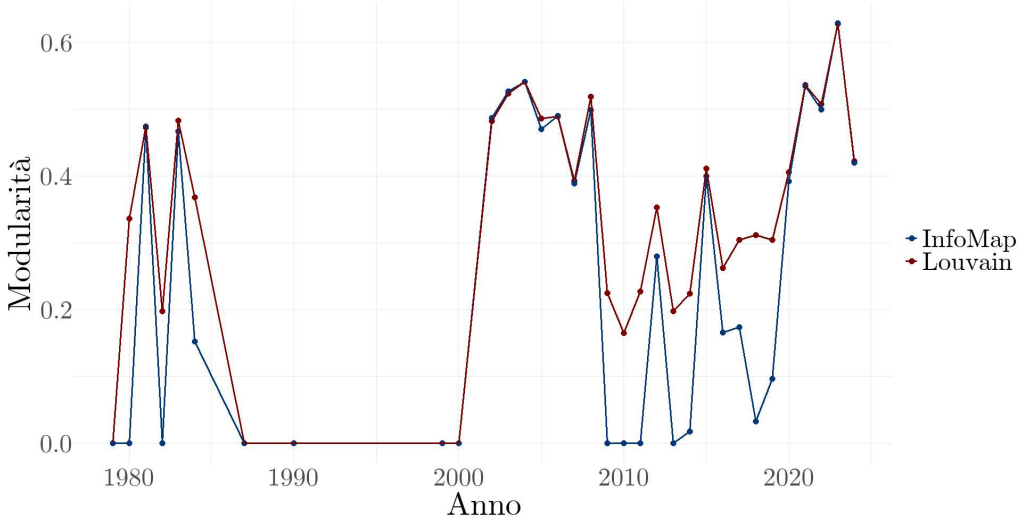


Figura 3.2: Andamento della modularità delle partizioni ottenute con l'algoritmo di Louvain e *InfoMap* nelle reti maschili.

Si consideri la rete del 2023, nella quale i due metodi portano a ottenere delle partizioni molto simili in termini di modularità. Louvain ha rilevato 10 comunità, mentre *InfoMap* ne ha identificato 12 (Figura 3.3). Un elemento chiave da osservare è la difficoltà nell'individuazione delle comunità di atleti che hanno partecipato a più competizioni, rappresentate nella parte centrale della rete. Questi atleti mostrano un maggior numero di connessioni con nodi in molte comunità, rendendo più difficile la loro assegnazione a una di esse. Tale fenomeno si osserva in entrambe le partizioni ma *InfoMap* sembra essere leggermente più sensibile, rilevando alcune comunità più piccole rispetto a Louvain. D'altra parte, le comunità periferiche, ossia quel-

le formate da atleti che hanno gareggiato solo una volta, sono state correttamente identificate da entrambi gli algoritmi.

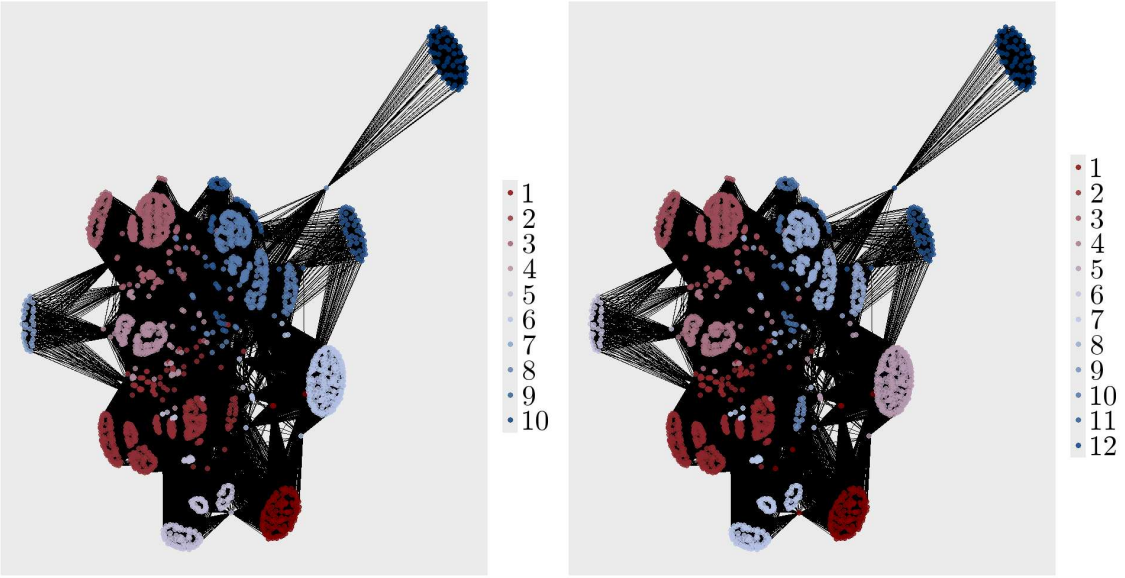


Figura 3.3: Rete maschile del 2023: comunità rilevate dagli algoritmi di Louvain (a sinistra) e *InfoMap* (a destra).

A partire dalla partizione ottenuta mediante l'algoritmo di Louvain, si valutano le caratteristiche degli atleti nelle 10 comunità rilevate in termini di peso corporeo e di prestazioni. È stato condotto un *test* ANOVA per verificare le differenze tra le comunità. Non si osserva una differenza significativa nel peso corporeo (*p-value* pari a 0.6) e nel numero di tentativi falliti (*p-value* pari a 0.2). Tuttavia, emerge una differenza significativa per il totale sollevato, con un *p-value* inferiore a 0.001. Focalizzandosi su quest'ultima variabile, nella Tabella 3.1 si nota che il valore medio massimo è attribuito alla decima comunità, la quale comprende atleti che registrano, in media, anche il peso corporeo più elevato. Interessante è anche la comunità 5, dove, nonostante il peso corporeo ridotto degli atleti, si osserva uno dei valori più elevati del totale sollevato. Questo gruppo include principalmente atleti della categoria più giovane (*sub-junior*), risultando quindi particolarmente competitivo.

Comunità	n	Peso corporeo	Totale sollevato	*Squat*	*Bench*	*Deadlift*	Tentativi falliti
1	112	84.55	328.97	85.76	90.07	149.80	1
2	601	84.28	424.70	129.48	94.13	206.37	2
3	785	86.01	431.12	123.26	92.23	218.80	2
4	296	86.71	191.61	21.31	123.81	49.45	1
5	203	77.57	437.91	156.52	97.27	189.84	2
6	377	87.24	183.53	46.06	53.58	83.88	1
7	31	86.32	361.69	118.87	108.87	144.03	1
8	423	82.99	405.13	137.94	112.15	167.73	2
9	110	85.88	335.18	60.34	112.52	69.02	1
10	68	93.91	451.50	145.76	116.19	189.54	2

Tabella 3.1: Caratteristiche degli atleti del 2023 nelle comunità identificate dall'algoritmo di Louvain: numerosità e media del peso corporeo, dei carichi sollevati e dei tentativi falliti.

Si procede in modo analogo per le reti nel contesto femminile, applicando anche l'algoritmo Girvan-Newman. Tuttavia, a partire dal 2017, l'aumento delle dimensioni delle reti ha comportato un significativo prolungamento dei tempi di calcolo per l'identificazione delle comunità, rendendo meno praticabile l'impiego di quest'ultimo algoritmo.

L'andamento della modularità, illustrato nella Figura 3.4, evidenzia che, come per le reti maschili, l'algoritmo di Louvain fornisce generalmente partizioni di qualità superiore. Si può notare che, nonostante i tempi di esecuzione più lunghi, l'algoritmo di Girvan e Newman ha prodotto partizioni con valori di modularità generalmente superiori a quelli di *InfoMap* e spesso comparabili a quelli delle configurazioni generate dall'algoritmo di Louvain. La limitata partecipazione delle atlete e la scarsità di competizioni nel primo periodo hanno reso più complessa l'identificazione delle comunità, in particolare utilizzando il metodo *InfoMap*, le cui partizioni mostrano infatti valori inferiori rispetto agli altri due algoritmi. Inoltre, come osservato per le reti degli atleti, i picchi negativi o i valori nulli della modularità si riscontrano in corrispondenza delle reti caratterizzate dalla presenza di atleti

con alti valori di *betweenness*. Fatta eccezione per l'algoritmo Girvan-Newman, la ridotta numerosità generale consente di ottenere tempi di esecuzione più rapidi, mettendo in evidenza una maggiore efficienza del metodo Louvain.

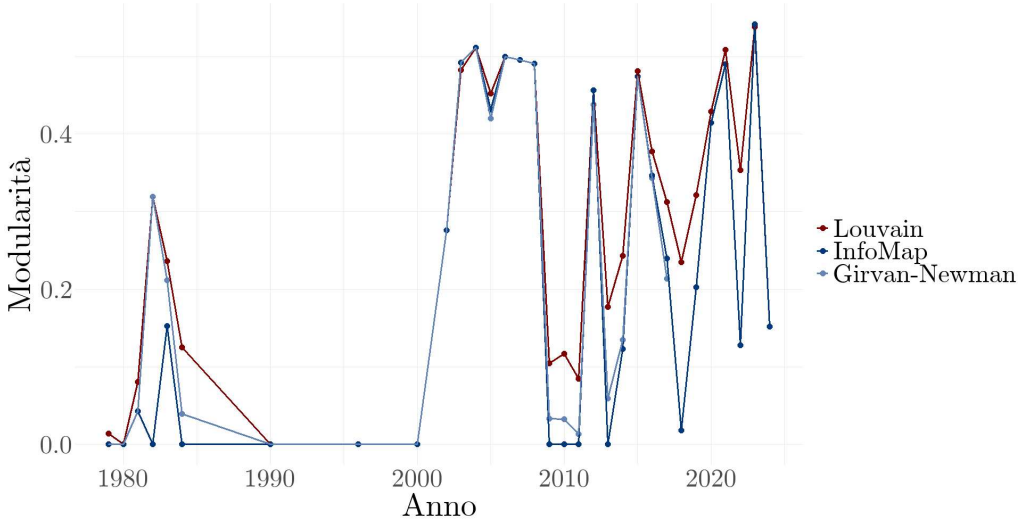


Figura 3.4: Andamento della modularità delle partizioni ottenute con l'algoritmo di Louvain e *InfoMap* nelle reti femminili.

Nella rete del 2023 sia l'algoritmo di Louvain che *InfoMap* portano ad identificare 10 comunità (Figura 3.5) portando ad un valore di modularità pari all'incirca a 0.54. Le poche differenze si concentrano principalmente sul gruppo di atlete nella componente maggiore della rete, caratterizzata da un maggior numero di connessioni.

Considerando le caratteristiche e le prestazioni delle atlete, sono emerse differenze significative sia nel totale sollevato che nel numero di tentativi falliti, con *p-value* rispettivamente pari a 0.002 e inferiore a 0.001. Al contrario, le differenze di peso corporeo tra le comunità non risultano rilevanti, similmente a quanto osservato per le comunità di atleti maschili.

A partire dalle caratteristiche delle atlete appartenenti alle comunità identificate dall'algoritmo di Louvain riportate nella Tabella 3.1, un aspetto interessante da evidenziare è che, come nel caso del decimo gruppo nelle reti maschili, anche la nona comunità femminile raggruppa atlete con un peso corporeo e un totale sollevato superiore. La seconda comunità si distingue invece come la più competitiva, evidenziando uno dei valori più elevati del totale sollevato e, allo stesso tempo, un peso corporeo relativamente contenuto. Entrambi questi gruppi sono prevalentemente composti da atlete appartenenti alla categoria *open*.

Per quanto riguarda la decima comunità, il valore nullo per la media del totale sollevato nello *squat* potrebbe inizialmente far pensare che includa solamente atlete che hanno partecipato a competizioni incomplete. Tuttavia, questa comunità raggruppa i casi in cui non è stato specificato il risultato delle singole prove, ma solo il totale complessivamente sollevato. I valori per le distensioni su panca e lo stacco da terra non sono pari a zero grazie alla presenza di atlete che hanno preso parte a competizioni con una o al massimo due alzate.

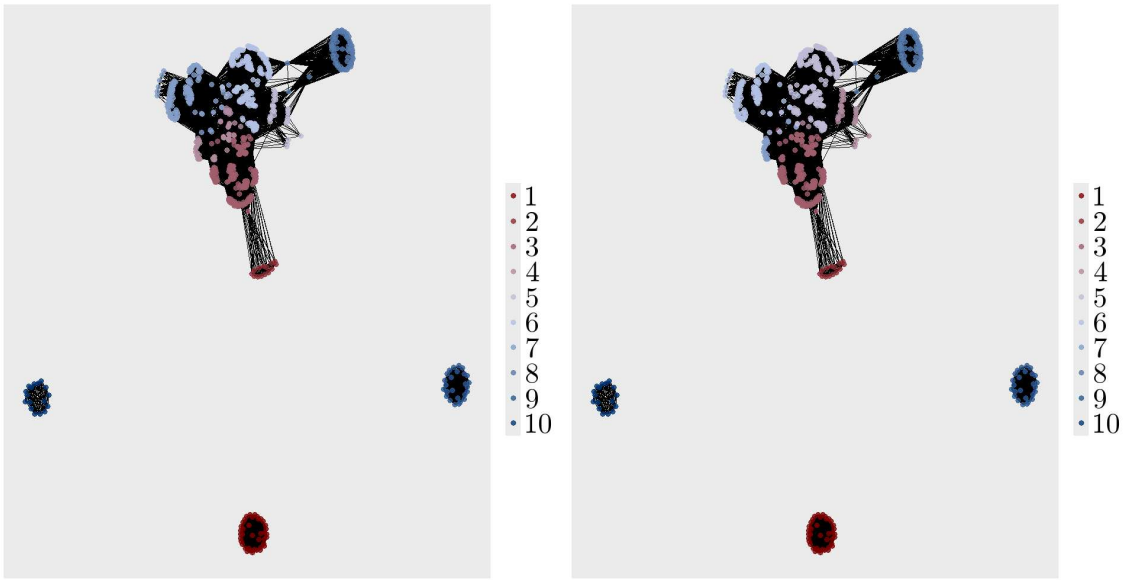


Figura 3.5: Rete femminile del 2023: comunità rilevate dagli algoritmi di Louvain (a sinistra) e *InfoMap* (a destra).

Comunità	n	Peso corporeo	Totale sollevato	*Squat*	*Bench*	*Deadlift*	Tentativi falliti
1	48	60.85	196.35	57.03	37.29	102.03	1
2	14	62.29	296.00	109.32	57.18	129.50	2
3	355	61.97	235.65	74.76	49.60	118.47	2
4	113	59.48	202.83	53.87	50.53	98.43	2
5	47	60.44	193.65	52.82	68.90	75.12	2
6	466	62.98	222.78	56.50	49.06	119.15	2
7	185	60.11	245.35	86.75	55.04	107.98	2
8	133	62.01	98.93	25.36	21.93	51.65	1
9	41	68.63	299.34	108.05	63.37	137.83	2
10	17	59.95	228.82	0.00	5.74	17.06	0

Tabella 3.2: Caratteristiche delle atlete del 2023 nelle comunità identificate: numerosità e media del peso corporeo, dei carichi sollevati e dei tentativi falliti.

3.3.3 *Community detection* dinamica

Per effettuare il rilevamento delle comunità dinamico è stato utilizzato il pacchetto `DynComm`, che, tramite una versione modificata del metodo di Louvain, identifica le comunità all'interno della rete aggiungendo e rimuovendo nodi e archi con l'evolversi del tempo (Sarmento et al., 2019). A tal fine, sono state realizzate due funzioni. La prima ha l'obiettivo di creare una matrice delle connessioni tra coppie di atleti per un determinato anno. Inizialmente, vengono generate tutte le possibili combinazioni di coppie di nodi, escludendo quelle in cui gli atleti coincidono. Successivamente, per garantire la non direzionalità delle relazioni, vengono eliminate le righe che rappresentano la stessa coppia di nodi in ordine inverso. Una volta identificate le coppie di atleti che hanno partecipato insieme ad almeno una competizione, vengono mantenute solo le righe che presentano una connessione. La matrice così realizzata presenterà quattro colonne: le prime due sono relative alle coppie di nodi, la terza, necessaria principalmente per la fase di aggiornamento, indica la presenza di connessione mentre la quarta riporta l'anno considerato.

La seconda funzione è dedicata all'aggiornamento delle connessioni da un anno al successivo, verificando la presenza di archi comuni. Il processo si articola in diversi passaggi. Dopo aver creato le matrici delle connessioni per due anni successivi ($t-1$ e t), queste vengono confrontate per valutare la presenza di connessioni che permangono. Per le coppie di atleti non più presenti o che non hanno gareggiato assieme nell'anno corrente, il valore della terza colonna della matrice al tempo $t-1$ verrà impostato a 0, per indicare che i nodi e l'arco devono essere rimossi. Le coppie che, invece, continuano a presentare una connessione verranno mantenute solamente nella matrice dell'anno t con valore 1 nella terza colonna e vengono eliminate dalla matrice al tempo $t-1$ per evitare una successiva duplicazione. La matrice aggiornata sarà quindi data dall'unione delle due matrici con le modifiche appena descritte. La quarta colonna, relativa all'anno, è necessaria per eliminare man mano le righe relative agli anni precedenti a t e $t-1$.

Nell'analisi, è stato necessario limitare l'attenzione esclusivamente alle reti femminili dei primi anni disponibili, dal 1979 al 2010. Questo perché, con un numero di nodi significativamente più elevato, come quello che caratterizza le reti maschili, il costo computazionale aumenta drasticamente, impedendo di ottenere risultati per i singoli anni e, di conseguenza, per l'evoluzione temporale. Nonostante la minor numerosità nell'ambito del *Powerlifting* femminile nei primi anni, l'ultima iterazione per il rilevamento delle comunità nel 2010 ha richiesto all'incirca 68 minuti. Inoltre, si è continuato a considerare connessioni binarie (presenza di almeno una competizione in comune). Inizialmente le due funzioni appena descritte sono state costruite per considerare il numero di competizioni in comune come forza della connessione presente tra una coppia di atleti. Questo approccio non è stato poi utilizzato a causa dell'eccessivo tempo di esecuzione richiesto dall'algoritmo.

Si considerino quindi le reti di partecipazione competitiva femmi-

nile nel periodo tra il 1979 e il 2010. Nei primi anni la qualità delle partizioni, determinata dalla modularità, risulta ridotta probabilmente a causa della limitata presenza di atlete. Si osserva un primo picco nel 1982, la cui rete riportata in Figura 2.1 mostra due gruppi sufficientemente distinti, uniti da un singolo nodo. La struttura più chiara di questa rete ha dunque portato a ottenere una buona partizione. Nel primo periodo, inoltre, i gruppi di atlete che hanno partecipato a competizioni risultano molto differenti, non mostrando quasi nessuna connessione che permanga nel corso di questi anni. A partire dal 2000, la maggiore partecipazione e il numero più elevato di archi comuni ha portato a un miglioramento della modularità, fino a raggiungere il valore massimo di 0.51 nel 2008. Dal 2008 al 2009, si registra un calo del numero di partecipazioni comuni, passando da 145, registrato tra il 2007 e il 2008, a 92, il che potrebbe aver contribuito alla diminuzione della modularità osservata. Infine, la decrescita persiste anche dal 2009 al 2010. Le reti di questi due anni risultano molto dense, rendendo più complesso identificare comunità ben separate.

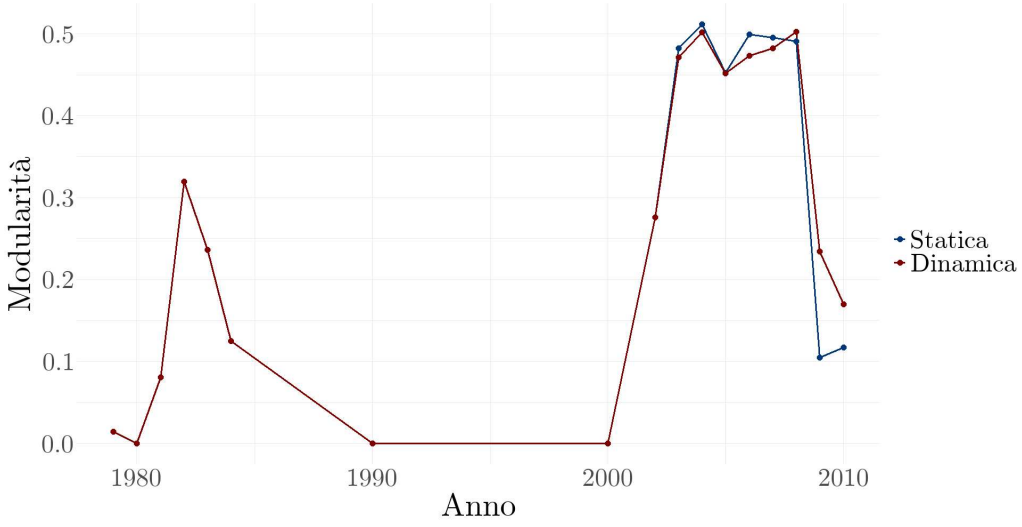


Figura 3.6: Rilevamento delle comunità statico e dinamico: andamento della modularità nelle reti femminili.

I valori di modularità delle partizioni ottenute attraverso le due versioni del rilevamento delle comunità tramite l'algoritmo di Louvain mostrano differenze solo a partire dal 2002. Da quell'anno, la *community detection* statica presenta valori leggermente più elevati, eccetto nel 2010. Tuttavia, il miglioramento riscontrato in quell'anno, che raggiunge un incremento solo di 0.05, non sembra giustificare l'adozione di questo approccio, soprattutto considerando il maggiore tempo di esecuzione richiesto.

Capitolo 4

Modelli per reti

La maggior parte delle reti mostra relazioni tra coppie di nodi. In questo elaborato viene considerata la connessione tra due atleti mediante la medesima partecipazione ad almeno una competizione in un determinato anno. Le coppie di nodi presenti in una rete vengono chiamate diadi e una quantità misurata su di esse costituisce una variabile diadica. La matrice di adiacenza, sia essa binaria o ponderata, riassume l'informazione fornita da una variabile diadica misurata su un insieme di nodi. In questo contesto, la matrice di adiacenza prende il nome di sociomatrice ($\mathbf{Y}$). I nodi posti in riga e in colonna nella sociomatrice vengono chiamati rispettivamente mittenti e riceventi. Così come per la matrice di adiacenza, se $y_{ij} = y_{ji}$ allora $\mathbf{Y}$ è simmetrica e la variabile diadica è indiretta.

La sociomatrice non solo sintetizza le relazioni tra diadi ma può anche fungere da base per la costruzione di modelli statistici. In questo capitolo verranno dunque esplorate diverse metodologie di modellistica utilizzate a partire dalle informazioni contenute nella sociomatrice. Si partirà dalla scomposizione ANOVA e da un modello ad effetti casuali definito a partire da quest'ultima, il *Social Relation Model* (SRM). Si passerà al *Social Relation Regression Model*, un'estensione del SRM per includere covariate all'interno del modello. Tuttavia, queste prime due soluzioni non sono sufficienti per considerare la transitività spesso presente nelle reti. Si introdurranno pertanto i modelli di rete

con effetti additivi e moltiplicativi (*Additive and Multiplicative Effects Network Model*, AME). La definizione di quest'ultima classe di modelli è, nella sua versione originale, adatta a variabili di natura continua e per sociomatrici asimmetriche. In molte applicazioni, tra cui quella di questo elaborato, sono però presenti connessioni binarie simmetriche. Il modello AME può dunque essere esteso per considerare entrambi questi aspetti.

Il capitolo si concluderà applicando tali modelli alle reti di partecipazione competitiva nel *Powerlifting*, con l'obiettivo di comprendere se le variabili disponibili a livello di nodo influenzano la formazione delle connessioni tra coppie di atleti, catturando al contempo effetti additivi e moltiplicativi in modo da tener conto anche di caratteristiche latenti non osservate.

La parte teorica di questo capitolo è basata sul lavoro di Hoff (2018).

4.1 ANOVA, SRM e SRRM

È usuale riscontrare una correlazione tra i valori di una variabile diadica in una specifica riga, il che implica che i valori alti e bassi non si distribuiscono uniformemente. Questo si traduce nella variabilità delle medie di riga di $\mathbf{Y}$, giustificabile dal fatto che i valori all'interno di una riga della sociomatrice riguardano lo stesso mittente. Ci si aspetta che se il mittente i risulta più "socievole" di j allora la media dell'i-esima riga sia maggiore di quella della riga j. Questo fenomeno viene definito come eterogeneità nella socialità. D'altra parte la variabilità osservata nelle medie di colonna viene invece identificata come eterogeneità nella popolarità.

Il modello più semplice per analizzare le due forme di eterogeneità si basa sulla scomposizione ANOVA (Hoff, 2018). In base a questo

modello, la variabile diadica y_{ij} viene espressa come

$$y_{ij} = \mu + a_i + b_j + \varepsilon_{ij}, \tag{4.1}$$

dove μ è la media generale, a_i e b_j corrispondono rispettivamente agli effetti di riga e di colonna, i quali riflettono l'eterogeneità nella socialità e nella popolarità, e ε_{ij} è la componente residuale.

È importante osservare che l'insieme di nodi posti in riga e in colonna all'interno della sociomatrice coincidono. In altre parole, un nodo i è al contempo mittente e ricevente, presentando così sia un effetto di riga a_i che di colonna b_i. È allora lecito supporre che i vettori $(a_1, \ldots, a_n)$ e $(b_1, \ldots, b_n)$ siano correlati. Inoltre, generalmente $y_{ij} \neq y_{ji}$ ed è quindi ragionevole supporre che anche le corrispettive componenti residuali, ε_{ij} e ε_{ji}, siano correlate.

Sia la relazione tra effetti di riga e di colonna che quella tra i residui ε_{ij} e ε_{ji}, nota come correlazione diadica, non vengono quantificate dal modello in 4.1. Viene di conseguenza proposto il *Social Relation Model* (SRM), un modello a effetti casuali definito a partire da 4.1 ma aggiungendo due assunzioni:

$$\begin{aligned} \{(a_1, b_1), \ldots, (a_n, b_n)\} &\overset{\text{i.i.d.}}{\sim} N_2(\mathbf{0}, \Sigma_{ab}), \\ \{(\varepsilon_{ij}, \varepsilon_{ji}) : i \neq j\} &\overset{\text{i.i.d.}}{\sim} N_2(\mathbf{0}, \Sigma_{\varepsilon}), \end{aligned} \tag{4.2}$$

dove

$$\Sigma_{a,b} = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix}, \quad \Sigma_{\varepsilon} = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

in cui σ_a^2 e σ_b^2 rappresentano rispettivamente l'eterogeneità di riga (socialità) e di colonna (popolarità), σ_{ab} riflette l'associazione lineare presente tra questi due effetti e ρ la correlazione diadica.

Per valutare l'associazione tra una variabile diadica e altre variabili a livello di nodo o diade è possibile utilizzare una combinazione del modello di regressione lineare e della struttura ipotizzata dal SRM.

Questo nuovo modello, chiamato *Social Relation Regression Model* (SRRM) (Hoff, 2018) esprime y_{ij} come

$$y_{ij} = \boldsymbol{\beta}^T \mathbf{x}_{ij} + a_i + b_j + \varepsilon_{ij},$$

dove $\boldsymbol{\beta}$ è il vettore di coefficienti di dimensione p associato alle variabili esplicative $\mathbf{x}_{ij}$ che può includere variabili nodali e/o diadiche.

4.2 *Additive and Multiplicative Effects Model* (AME)

La tendenza alla formazione di triangoli di nodi spesso riscontrata all'interno di una rete viene misurata dalla transitività. Questa dipendenza di terzo ordine, chiamata triadica, è quantificabile da statistiche riassuntive del tipo

$$\sum_{i \neq j \neq k} \frac{\hat{e}_{ij}\hat{e}_{ik}\hat{e}_{jk}}{n(n-1)(n-2)}, \tag{4.3}$$

dove gli $\hat{e}_{ij}$ indicano i residui del modello stimato mediante i minimi quadrati ordinari. Il modello SRRM non coglie la dipendenza triadica poiché, dato che $\hat{e}_{ij} \approx \hat{a}_i + \hat{b}_i + \varepsilon_{ij}$ e $\hat{a}_i$, $\hat{b}_i$ e ε_{ij} sono variabili casuali a media nulla, anche la quantità in 4.3 risulterà vicina a zero.

Per considerare questi schemi di ordine superiore una soluzione è quella di inserire un effetto moltiplicativo. Il modello a effetti additivi e moltiplicativi (AME) (Hoff, 2018) è quindi definito come

$$y_{ij} = \boldsymbol{\beta}^T \mathbf{x}_{ij} + a_i + b_j + \mathbf{u}_i^T \mathbf{v}_j + \varepsilon_{ij}, \tag{4.4}$$

dove $\mathbf{u}_i$ e $\mathbf{v}_j$ sono vettori di dimensione r rappresentanti rispettivamente le caratteristiche latenti del nodo i come mittente e del nodo j come ricevente. Il modello, oltre a 4.2, assume che

$$\{(\mathbf{u}_1, \mathbf{v}_1), \ldots, (\mathbf{u}_n, \mathbf{v}_n)\} \overset{\text{i.i.d.}}{\sim} N_{2r}(\mathbf{0}, \Psi).$$

Gli effetti moltiplicativi consentono di tener conto della dipendenza triadica presente all'interno della sociomatrice e possono rappresentare attributi a livello di nodo omesse.

Il modello in 4.4 viene chiamato AME Gaussiano poiché i dati, condizionatamente al vettore di parametri e agli effetti additivi e moltiplicativi, si distribuisce come una Normale.

Nel caso di reti indirette non è necessario distinguere gli effetti di mittente e ricevente. Il modello AME indiretto (Hoff, 2018) risulta

$$y_{ij} = \boldsymbol{\beta}^T \mathbf{x}_{ij} + \mathbf{u}_i^T \Lambda \mathbf{u}_j + a_i + a_j + \varepsilon_{ij},$$

dove $\{\varepsilon_{ij} : 1 \leq i > j \leq n\} \overset{\text{i.i.d.}}{\sim} N_2(\mathbf{0}, \Sigma_\varepsilon)$ e il termine moltiplicativo include ora come parametro una matrice diagonale Λ di "autovalori".

4.2.1 Modello AME per dati binari e ordinali

Variabili diadiche binarie o ordinali non possono essere rappresentate adeguatamente da un modello AME con termine di errore Gaussiano. In questi casi, si necessita di estendere il modello in 4.4 per tener conto della natura della connessione presente tra una coppia di nodi.

Sia $\mathbf{S}$ una sociomatrice per una determinata variabile diadica s_{ij}. Il caso più semplice di una variabile diadica ordinale è rappresentata da una variabile binaria $s_{ij} \in \{0, 1\}$ che indica se è presente o meno un arco tra una diade. Il modello di regressione *probit* esprime la probabilità della connessione tra i e j come $\Phi(\boldsymbol{\beta}^T \mathbf{x}_{ij})$, dove $\Phi(\cdot)$ è la funzione di ripartizione di una $N(0, 1)$. Questo modello ha una rappresentazione a variabile latente in cui s_{ij} è l'indicatore binario di una qualche variabile latente normale, $y_{ij} \sim N(\boldsymbol{\beta}^T \mathbf{x}_{ij}, 1)$. Tuttavia, l'assunzione alla base del modello che prevede l'indipendenza delle y_{ij} è inappropriata nel contesto delle reti, oltre al fatto che non viene colta alcun tipo di dipendenza.

Il modello AME per la variabile latente y_{ij} (Hoff, 2018) viene definito come

$$\begin{aligned} y_{ij} &= \boldsymbol{\beta}^T \mathbf{x}_{ij} + \mathbf{u}_i^T \mathbf{v}_j + a_i + b_j + \varepsilon_{ij} \\ s_{ij} &= g(y_{ij}) \end{aligned} \tag{4.5}$$

dove l'unica differenza rispetto a 4.4 è la presenza della funzione indicatrice $g(y) = \mathrm{I}(y > 0)$.

La naturale estensione del modello 4.5 alla presenza di una variabile ordinale con più di due livelli prevede di considerare una funzione $g(\cdot)$ che sia non decrescente.

4.3 Applicazione ai dati

La variabile diadica y_{ij} nelle reti delle partecipazioni competitive nel *Powerlifting* è binaria: assume valore 1 se gli atleti i e j hanno partecipato almeno una volta alla medesima competizione in un determinato anno e varrà 0 altrimenti. Essendo la connessione non direzionale, la sociomatrice risultante è simmetrica, quindi l'eterogeneità di riga e colonna coincidono e la dipendenza diadica sarà pari a 1.

Sebbene esista una versione del modello AME per dati longitudinali, in questo elaborato non è stata utilizzata. Tale approccio richiederebbe infatti che l'insieme dei nodi rimanga costante durante il periodo di osservazione, mentre nelle reti esaminate generalmente gli atleti non competono ogni anno. Questo porta a una presenza discontinua dei nodi, con atleti che possono partecipare in determinati anni, risultare assenti in altri e riapparire successivamente.

Prima di stimare i modelli è stata creata per ogni anno una matrice contenente le caratteristiche a livello di nodo. Le variabili considerate sono:

- la media del peso corporeo (`BodyweightKg`) nel caso di atleti che hanno partecipato a più competizioni in un determinato anno o il valore puntuale per chi ha effettuato una sola gara;
- la media del totale sollevato (`TotalKg`) o il valore puntuale;
- la media del numero di tentativi falliti (`Fail`) o il valore puntuale.

Le altre variabili rilevanti nell'insieme di dati, in parte analizzate nell'analisi esplorativa nella sezione 1.3, includono la/e tipologia/e di evento competitivo a cui l'atleta ha partecipato, l'equipaggiamento utilizzato nella competizione e la/e categoria/e di età in cui ha gareggiato. A partire da queste sono state create un numero di variabili indicatrici pari alle modalità che la variabile originaria presenta meno uno. La categoria *baseline* viene determinata ogni anno, selezionando quella più frequente. È importante notare che i livelli delle variabili categoriali variano nel tempo: ad esempio, alcune tipologie di competizioni possono essere presenti in certi anni e non in altri. Di conseguenza, le variabili *dummy* generate non saranno le stesse con l'evolversi del tempo e riflettono le modalità effettivamente presenti per ciascun anno.

L'obiettivo principale è comprendere se e come le caratteristiche dei nodi influenzano la formazione delle connessioni tra gli atleti. L'analisi si concentrerà prevalentemente sulle reti femminili, poiché la dimensione più ridotta delle loro sociomatrici rende la stima più gestibile in termini di tempi computazionali. Si partirà dal modello basato sulla scomposizione ANOVA per verificare l'eterogeneità tra gli atleti, per poi proseguire con le metodologie più complesse.

Rispetto all'approccio basato sulla scomposizione ANOVA, il modello SRM introduce la relazione tra effetti di riga e colonna e la dipendenza diadica. Tuttavia, la simmetria della sociomatrice elimina la necessità di distinguere tra mittenti e riceventi, mentre la dipendenza diadica rimane sempre pari a 1. Pertanto, i risultati di questa metodologia non verranno riportati.

4.3.1 ANOVA e SRRM

Si consideri la rete delle partecipazioni competitive delle atlete nel 2023. Nel modello basato sulla scomposizione ANOVA, la presenza o assenza di connessione tra una coppia di atlete è descritta solamente dalla media complessiva e dall'effetto di riga o, equivalentemente, di colonna. I risultati riportati in Tabella 4.1 mostrano un effetto di riga altamente significativo, indicando una forte eterogeneità tra le atlete nella formazione di connessioni. In altre parole, alcune atlete presentano un numero significativamente superiore di connessioni rispetto ad altre, il che è indicativo di una maggiore partecipazione alle competizioni.

	Df	Sum Sq	Mean Sq	F value	$\Pr(>F)$
Row.athlete	754	4291	5.6904	47.882	<2.2e-16
Residuals	569270	67652	0.1188		

Tabella 4.1: Risultati del modello ANOVA per la rete femminile del 2023.

Le stime dei modelli SRRM e AME vengono ottenute in `R` mediante le funzioni della libreria `amen` (Hoff et al., 2024). I parametri ignoti dei modelli sono stimati tramite un algoritmo di campionamento iterativo. Questo processo utilizza una catena di Markov per approssimare la distribuzione a posteriori dei parametri di interesse, partendo da distribuzioni a priori definite. L'algoritmo adotta un *Gibbs sampler* che, iterando, simula ogni parametro dalla sua distribuzione condizionale, dati i valori attuali degli altri parametri e i dati disponibili. Per la valutazione del modello si utilizzano riepiloghi delle statistiche della rete basati sulla distribuzione predittiva a posteriori. In particolare queste misure sono: la deviazione standard empirica delle medie di riga e di colonna, la correlazione diadica e una versione normalizzata della dipendenza triadica. Nella pratica, la valutazione della bontà di adattamento avviene confrontando i valori simulati dalla distribu-

zione predittiva a posteriori, rappresentati mediante istogrammi, con le statistiche osservate nella rete reale, indicate da linee verticali nei grafici. La sovrapposizione di queste ultime ai rispettivi istogrammi indica una buona capacità del modello di rappresentare la struttura della rete (Hoff, 2015).

La Figura 4.1, relativa al modello SRM stimato, mostra la distribuzione a posteriori della deviazione standard delle medie di riga (in alto), la distribuzione della dipendenza diadica (in basso a sinistra) e quella della dipendenza triadica (in basso a destra). Non verranno invece riportati gli istogrammi relativi agli effetti di colonna, che, come anticipato, coincidono con quelli di riga a causa della simmetria della sociomatrice.

Dai grafici si evidenzia che l'eterogeneità di riga non è adeguatamente catturata dal modello, il quale risulta quindi avere un adattamento insoddisfacente. Al contrario, non ci si aspetta che la dipendenza triadica venga ben rappresentata, poiché il modello non la considera.

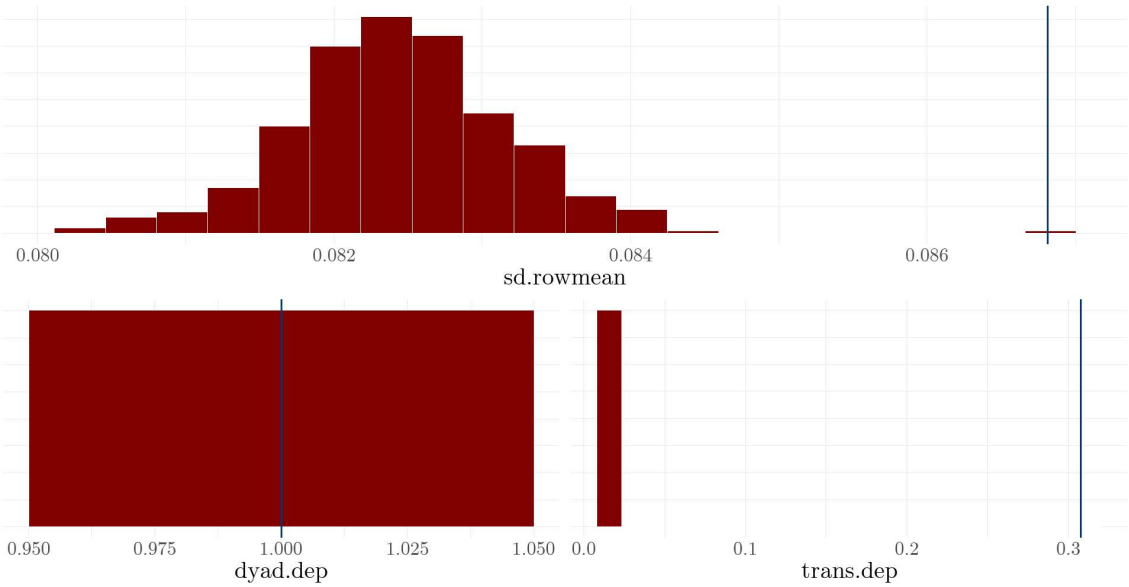


Figura 4.1: Bontà di adattamento del modello SRRM per la rete femminile del 2023.

	pmean	psd	z-stat	p-val
intercept	-2.160	0.152	-14.240	0.000
Event_B	0.437	0.028	15.417	0.000
Event_D	0.373	0.030	12.619	0.000
Event_BD	-0.355	0.324	-1.098	0.272
Event_S	-0.151	0.063	-2.404	0.016
Equipment_Wraps	-0.399	0.068	-5.906	0.000
Equipment_Single-ply	-0.320	0.042	-7.714	0.000
Division_Junior	0.124	0.032	3.825	0.000
Division_Master I	-0.111	0.041	-2.701	0.007
Division_Open	-0.167	0.028	-5.871	0.000
Division_Sub-Junior	0.086	0.036	2.405	0.016
Division_Master II	-0.124	0.058	-2.155	0.031
Division_Master III	0.016	0.070	0.236	0.814
Division_Master IV	-0.081	0.103	-0.790	0.430
BodyweightKg	-0.003	0.001	-3.042	0.002
TotalKg	0.002	0.000	11.070	0.000
Fail	0.059	0.010	5.610	0.000

Tabella 4.2: Stime dei coefficienti del modello SRRM per la rete femminile del 2023.

La Tabella 4.2 mostra, nell'ordine, le stime dei coefficienti del modello, ossia le medie a posteriori, le deviazioni standard a posteriori, gli *z-score* e i *p-value*. Si evidenzia che la maggior parte delle variabili è significativa. La variabile `Event`, eccetto per la modalità BD, ha un effetto rilevante sulla formazione delle connessioni. Il valore di riferimento per questa variabile è SBD (gare complete). Dai risultati emerge che la partecipazione ad eventi in cui viene eseguita la singola alzata di stacco da terra (D) o di distensioni su panca (B), favorisce maggiormente la formazione di connessioni tra le atlete, rispetto alla partecipazione a competizioni complete. Al contrario, partecipare a gare in cui è prevista la sola alzata dello *squat* riduce la probabilità di connessioni rispetto alle competizioni SBD.

L'equipaggiamento utilizzato (`Equipment`) è risultato significativo e mostra un effetto negativo rispetto alla modalità *raw* (equipaggiamento base), che è il valore di riferimento. In particolare, l'uso di fasce per le ginocchia (*wrap*) o di attrezzatura *single-ply* riduce la probabilità di formazione di connessioni tra le atlete rispetto ad utilizzare

l'equipaggiamento di base previsto nelle competizioni.

Le diverse categorie di età mostrano effetti distinti rispetto alla *baseline senior* e risultano significative eccetto nei due casi di *master* III e *master* IV. Le categorie *junior* e *sub-junior* presentano un effetto positivo. Al contrario, le categorie *master* I e *master* II hanno un effetto negativo, suggerendo che gli atleti più anziani tendono a formare meno connessioni rispetto alla categoria di riferimento.

Le ultime tre variabili sono relative al peso corporeo, al totale sollevato e al numero di tentativi falliti. Il peso corporeo mostra un'influenza negativa sulla variabile risposta, suggerendo che le atlete con un peso maggiore tendono ad avere una probabilità inferiore di formare connessioni. Questo effetto è coerente con l'effetto positivo riscontrato per la categoria *junior*, che include atlete più giovani e spesso di peso inferiore. Per quanto riguarda il totale sollevato durante le competizioni, il coefficiente positivo indica che atlete con prestazioni più alte tendono ad avere una maggiore propensione a stabilire connessioni. Questo riflette dunque una partecipazione più frequente alle competizioni per atlete di alto livello. Si specifica comunque che i coefficienti relativi a `BodyweightKg` e `TotalKg` sono vicini a zero, suggerendo che, pur essendo significativi, i loro effetti sulla variabile risposta sono relativamente contenuti. Infine, il numero di tentativi falliti (`Fail`) presenta un effetto positivo.

L'adattamento poco soddisfacente del modello potrebbe dipendere dalla presenza di tre componenti nella rete. Questa particolare configurazione di nodi isolati potrebbe ostacolare la capacità del modello di rappresentare adeguatamente le dinamiche delle connessioni tra le atlete. È quindi stato stimato lo stesso modello escludendo le atlete di tali componenti, corrispondenti alle ultime tre comunità identificate dall'algoritmo di Louvain nell'ambito della *community detection* (Figura 3.5). I grafici diagnostici evidenziano un leggero miglioramento rispetto al modello che includeva tutte le atlete, ma mostrano ancora un'incapacità di catturare pienamente l'eterogeneità presente.

4.3.2 Modello AME

Come evidenziato dalla formula del modello SRRM e dai grafici in Figura 4.1, questa metodologia non considera la dipendenza triadica. Il modello AME consente di tener conto non solo di questa forma di dipendenza, ma anche dell'eventuale presenza di variabili omesse o non osservabili che possono influenzare la formazione di connessioni. Risulta in questo caso necessario determinare la dimensione r del vettore delle caratteristiche latenti; nel primo modello AME stimato, $r = 2$.

Dai grafici in Figura 4.2 si nota un miglioramento dell'adattamento rispetto al modello SRRM. Viene catturata l'eterogeneità delle atlete, anche se la distribuzione non appare centrata sulla deviazione standard delle medie di riga della sociomatrice (σ_a). La discrepanza evidenziata nel grafo in basso a destra indica che il modello non spiega adeguatamente la dipendenza triadica. Infatti, si ha che $\hat{\rho} = 0.26$ quando il valore osservato è invece $\rho = 0.31$.

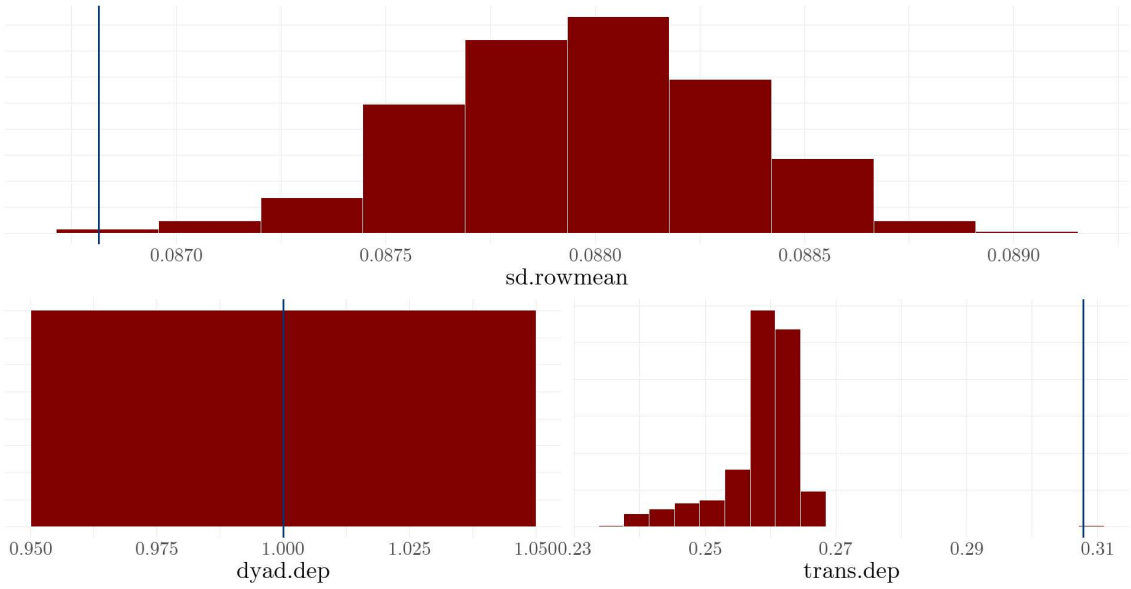


Figura 4.2: Bontà di adattamento del modello AME con $r = 2$ per la rete femminile del 2023.

	pmean	psd	z-stat	p-val
Intercept	-10.452	1.804	-5.794	0.000
Event_B	2.125	0.420	5.056	0.000
Event_D	0.771	0.212	3.645	0.000
Event_BD	0.667	2.363	0.282	0.778
Event_S	-0.786	0.628	-1.252	0.211
Equipment_Wraps	-9.802	2.389	-4.103	0.000
Equipment_Single-ply	1.082	0.325	3.333	0.001
Division_Junior	2.254	0.649	3.473	0.001
Division_Master I	2.759	0.455	6.066	0.000
Division_Open	-0.320	0.938	-0.341	0.733
Division_Sub-Junior	1.270	0.512	2.480	0.013
Division_Master II	1.861	0.546	3.408	0.001
Division_Master III	1.860	0.590	3.153	0.002
Division_Master IV	-1.069	0.919	-1.163	0.245
BodyweightKg	-0.014	0.009	-1.438	0.151
TotalKg	0.002	0.001	2.122	0.034
Fail	0.140	0.093	1.506	0.132

Tabella 4.3: Stime dei coefficienti del modello AME con $r = 2$ per la rete femminile del 2023.

Dalla Tabella 4.3 emerge che la variabile `Event_S`, insieme a quelle relative al peso corporeo e al numero di tentativi falliti, perde di significatività rispetto al modello SRRM. Si può poi osservare che i valori assoluti dei coefficienti nel modello AME sono generalmente superiori rispetto a quelli riportati in Tabella 4.2. Inoltre, per le due variabili `Division_Master I` e `Division_Master II`, le stime presentano un cambiamento di segno, indicando un effetto positivo sulla variabile risposta rispetto alla categoria di riferimento della divisione.

Considerando l'unica variabile di natura numerica significativa nel modello AME, viene riportata in Figura 4.3 la rete in cui il colore dei nodi è proporzionale al totale sollevato. Osservando la componente principale, emerge una maggiore concentrazione di atlete con un totale sollevato superiore nella parte centrale della rete caratterizzata da un maggior numero di connessioni. Tuttavia, sono presenti anche nodi con totali sollevati ridotti. Ciò non implica necessariamente scarse prestazioni, poiché alcune di queste atlete hanno partecipato a competizioni con una sola alzata, come indicato dalla forma assegnata ai nodi. Sia

questo aspetto che la presenza di alcuni *hub* con un valore di `TotalKg` inferiore potrebbero giustificare il coefficiente ridotto della variabile, evidenziandone un effetto limitato sulla formazione di connessioni.

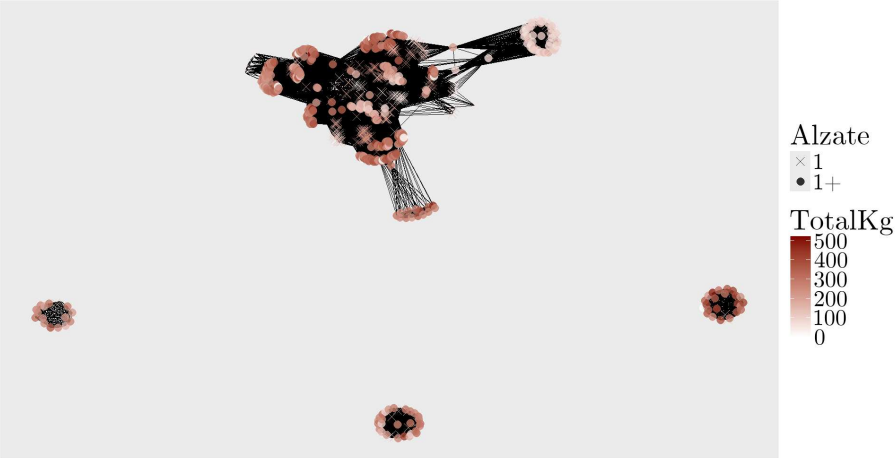


Figura 4.3: Rete femminile del 2023 differenziata per totale sollevato e numero di alzate svolte nelle competizioni.

Aumentando la dimensione del fattore latente da 2 a 10, si osserva un miglioramento nell'adattamento del modello sia in termini di eterogeneità di riga che di dipendenza triadica (Figura 4.4). Tuttavia, nonostante la minor distanza tra ρ e $\hat{\rho}$, che differiscono solamente di 0.003, il modello resta limitato nella capacità di descrivere accuratamente la dipendenza triadica.

I coefficienti delle variabili indicatrici relative alle divisioni *master* I, II, III così come quella relativa all'equipaggiamento *single-ply* perdono di significatività rispetto allo stesso modello con $r = 2$. Ritorna invece significativo il peso corporeo e il numero di tentativi falliti. Confrontando le tabelle 4.3 e 4.4 non si verificano cambiamenti di segno dei coefficienti. Si noti che quest'ultimo modello, caratterizzato da un miglior adattamento, prevede un effetto significativo solamente per le categorie di età più giovani, indicando una maggiore propensione alla formazione di connessioni di queste rispetto ad atlete di età superiore.

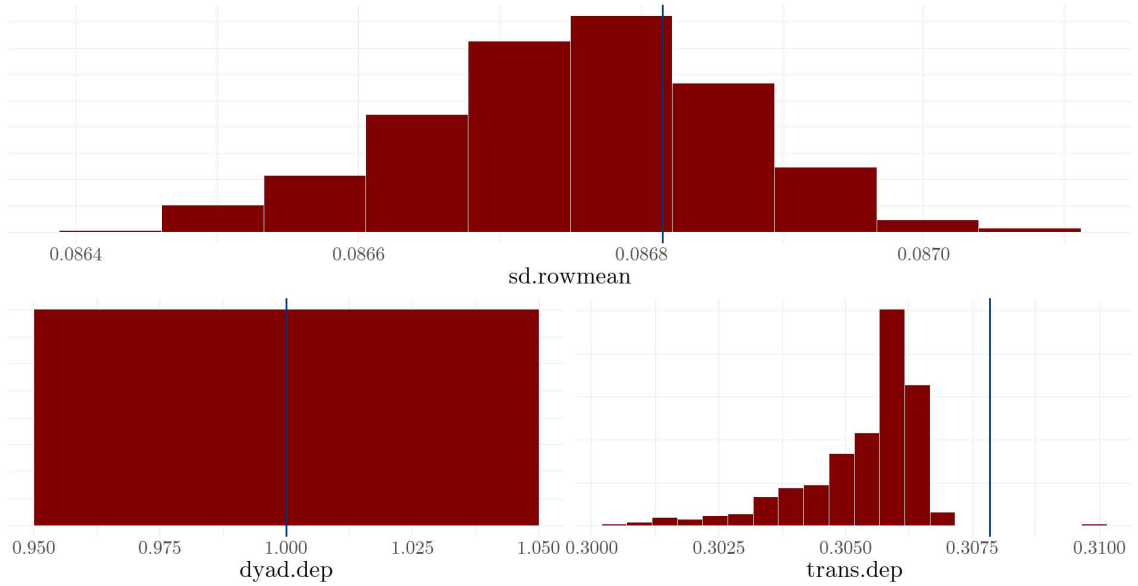


Figura 4.4: Bontà di adattamento del modello AME con $r = 10$ per la rete femminile dell 2023.

	pmean	psd	z-stat	p-val
Intercept	-11.897	2.224	-5.349	0.000
Event_B	1.757	0.377	4.665	0.000
Event_D	1.485	0.141	10.500	0.000
Event_BD	0.034	0.998	0.034	0.973
Event_S	-0.150	0.165	-0.909	0.364
Equipment_Wraps	-0.408	0.176	-2.318	0.020
Equipment_Single-ply	0.817	0.533	1.533	0.125
Division_Junior	0.466	0.097	4.818	0.000
Division_Master I	-0.103	0.116	-0.887	0.375
Division_Open	0.273	0.177	1.544	0.123
Division_Sub-Junior	0.331	0.153	2.171	0.030
Division_Master II	-0.082	0.223	-0.367	0.713
Division_Master III	0.199	0.178	1.115	0.265
Division_Master IV	-0.301	0.434	-0.692	0.489
BodyweightKg	-0.007	0.003	-2.134	0.033
TotalKg	0.005	0.001	8.127	0.000
Fail	0.124	0.061	2.020	0.043

Tabella 4.4: Stime dei coefficienti del modello AME con $r = 10$ per la rete femminile del 2023.

Nella Tabella 4.5 la grandezza degli elementi sulla diagonale di Λ, indica che tutti i fattori latenti considerati hanno un importante contributo sulla connessione tra atleti. Inoltre, il segno positivo suggerisce la presenza di omofilia. Poiché i fattori latenti rappresentano variabili non osservate, risulta complesso fornirne un'interpretazione. Tuttavia, osservando la correlazione tra i fattori latenti e le covariate, emerge come alcune variabili non significative potrebbero influire sulle connessioni in modo moltiplicativo piuttosto che additivo (Hoff, 2015).

k	λ_k
1	2904.01
2	2571.25
3	1862.88
4	1373.46
5	1040.01
6	899.45
7	827.55
8	716.80
9	559.04
10	437.95

Tabella 4.5: "Autovalori" $\lambda_k, k = 1, \ldots, 10$ stimati della matrice Λ.

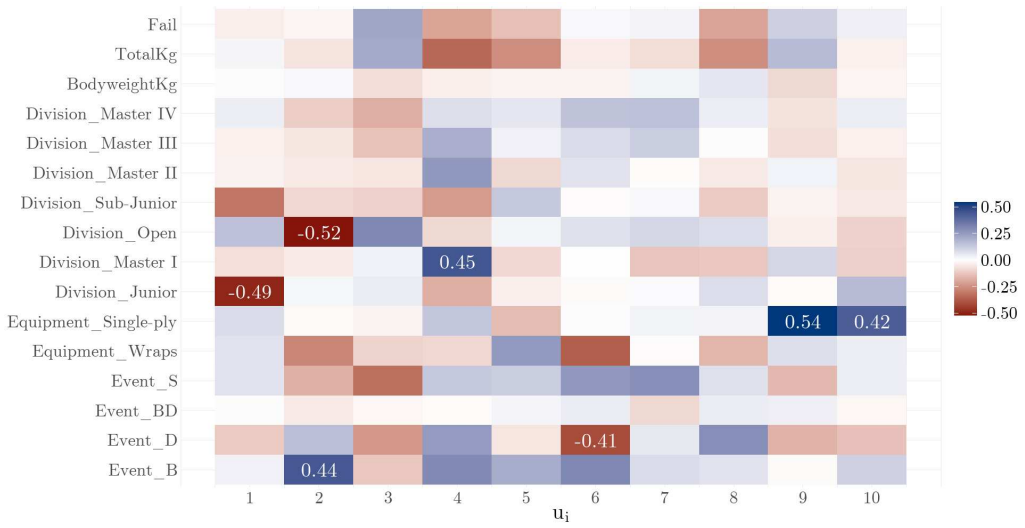


Figura 4.5: Correlazione tra il vettore di fattori latenti $\mathbf{u}_i$ e le variabili esplicative.

Anche per il modello AME si è tentato di escludere le atlete delle comunità isolate della rete, per verificare se questa particolare configurazione fosse responsabile dei risultati non completamente ottimali. Tuttavia, a differenza di quanto osservato nel modello SRRM, questa esclusione non ha portato a miglioramenti nella bontà di adattamento. Infatti, il modello ha continuato a mostrare difficoltà nel cogliere adeguatamente la dipendenza triadica.

Confrontando i risultati ottenuti in anni precedenti, caratterizzati da una minore numerosità, con quelli del 2023, emerge un adattamento significativamente migliore del modello nei dati degli anni passati. In questi casi, i grafici di diagnostica mostrano un'ottima sovrapposizione tra i valori osservati e quelli a posteriori delle statistiche riassuntive, indicando una maggiore capacità del modello di descrivere le relazioni tra le atlete nelle reti meno complesse. L'incremento della numerosità, oltre a causare un evidente peggioramento nell'adattamento, ha anche comportato un aumento sostanziale del tempo computazionale richiesto per la stima.

Considerando la rete maschile del 2023, la cui sociomatrice ha dimensione 1845×1845, la stima del modello AME si è rivelata particolarmente impegnativa, sia con $r = 2$ che, ancor più, con $r = 10$. Di conseguenza, è risultata necessaria una sensibile riduzione del numero di iterazioni della catena di Markov. Questa diminuzione ha compromesso i risultati del modello, il quale ha mostrato notevoli difficoltà nel catturare sia l'eterogeneità presente tra gli atleti che la dipendenza triadica. Per questi motivi si è deciso di non riportare i risultati ottenuti.

Conclusione

L'obiettivo principale di questo elaborato è stato quello di esplorare ed analizzare le dinamiche delle interazioni tra gli atleti di *Powerlifting* in Italia, intese come partecipazione ad almeno una stessa competizione. Queste reti sono state costruite a partire dall'insieme di dati rilasciato da *OpenPowerlifting* che costituisce l'unica raccolta attualmente presente delle prestazioni degli atleti registrate nelle competizioni di ogni livello in tutto il mondo. Poiché non esistono gare che vedano coinvolti sia uomini che donne, è risultata necessaria una separazione delle reti per genere.

Dopo aver introdotto lo sport del *Powerlifting*, descrivendone il regolamento, si è passati ad una breve analisi esplorativa dei dati, concentrandosi sulle variabili di maggiore rilevanza, riguardanti le caratteristiche degli atleti e degli eventi competitivi. Questa analisi iniziale ha fornito una visione complessiva del *dataset*, permettendo una migliore comprensione delle dinamiche di questo sport.

Il secondo capitolo è stato dedicato ai concetti fondamentali dell'analisi di dati di rete, illustrando le statistiche descrittive comunemente utilizzate e le proprietà maggiormente riscontrate. Una definizione equivalente della rete di partecipazione competitiva inizialmente descritta è fornita dalla rete bipartita nella quale si hanno due tipologie di nodi: gli atleti e le competizioni. La complessità interpretativa all'aumentare delle dimensioni ha portato a preferire la struttura di rete con un solo tipo di nodi, rappresentante gli atleti.

Grazie alla disponibilità di dati per diversi anni, è stato possibile un approccio dinamico alle reti, adottando la struttura delle reti tem-

porali. Utilizzando le statistiche descrittive proposte, si è osservato che le dimensioni delle reti restano relativamente contenute. La maggiore partecipazione degli atleti alle competizioni porta valori medi del grado uniformemente più elevati rispetto a quanto osservato per le reti femminili. Attraverso il grado e l'indice di centralità della *betweenness*, è stato inoltre possibile identificare atleti con un alto numero di connessioni e in posizioni strategiche all'interno della rete.

Si è visto come una disposizione casuale dei nodi porti a non preservare importanti caratteristiche della rete. L'utilizzo di algoritmi di posizionamento *force-directed* ha permesso invece di evidenziare facilmente elementi chiave come la centralità di un nodo o la presenza di comunità. In particolare, sono stati illustrati i due metodi più diffusi tra quelli appartenenti alla classe di algoritmi *fore-directed*: Fruchterman-Reingold e Kamada-Kawai. Per ragioni di efficienza computazionale, si è preferito il primo, il quale ha consentito una disposizione rapida dei nodi mantenendo una chiara identificazione delle caratteristiche proprie delle reti.

Il rilevamento delle comunità ha reso possibile l'identificazione di alcuni gruppi di atleti e atlete maggiormente competitivi. Qualora le loro prestazioni risultassero raggiungibili, l'identificazione di questi gruppi potrebbe aiutare i preparatori di atleti con caratteristiche simili a sviluppare programmi di allenamento mirati a superarli in eventuali future competizioni. I tre metodi proposti (Louvain, Girvan-Newman e *InfoMap*) hanno generalmente prodotto configurazioni simili, con Louvain che si è dimostrato il metodo migliore in termini di modularità. D'altra parte, i tempi richiesti dall'algoritmo di Girvan e Newman hanno impedito il suo utilizzo nelle reti maschili a causa delle loro elevate dimensioni. È stato poi considerato un approccio dinamico, per considerare l'apparizione e la rimozione di nodi nel corso degli anni. La difficoltà di implementazione al crescere delle dimensioni e i valori pressoché uguali della modularità hanno portato a preferire l'approccio statico rispetto alla rilevazione dinamica.

Sulla base delle variabili a livello di nodo e concentrandosi sulle reti femminili del 2023, sono stati stimati modelli per valutare come le caratteristiche delle atlete influenzino la formazione delle connessioni. Il modello ANOVA ha confermato la presenza di una significativa eterogeneità tra le atlete, mentre i modelli SRRM e AME hanno analizzato più specificamente l'effetto delle variabili. Si è ottenuto il miglior adattamento con il modello AME con $r = 10$, nonostante alcune difficoltà nel cogliere la dipendenza triadica. La partecipazione a competizioni focalizzate su una singola alzata, come distensioni su panca o stacco da terra, sembra favorire la formazione di connessioni, probabilmente per il minor sforzo richiesto rispetto alle gare SBD che potrebbe incentivare la partecipazione a più eventi. Le variabili relative alle categorie *sub-junior* e *junior* suggerisce una crescente partecipazione delle nuove generazioni. Infine, anche il peso corporeo e le prestazioni in gara mostrano un impatto significativo sulla formazione delle connessioni.

Il presente elaborato ha limitato l'analisi al contesto italiano, ma una possibile estensione futura potrebbe includere più Paesi, permettendo così un confronto internazionale delle reti competitive nel *Powerlifting*. Tuttavia, la numerosità dei dati potrebbe comportare problematiche legate ai tempi e all'efficienza computazionale, rendendo necessaria l'adozione di metodologie ottimizzate per reti di grandi dimensioni. Un ulteriore sviluppo, vantaggioso a livello strategico, riguarda l'utilizzo di modelli predittivi per prevedere la formazione di connessioni, al fine di valutare la probabilità che due atleti si incontrino nuovamente in una competizione futura.

Appendice A

Materiale aggiuntivo capitolo 1

A.1 Storia del *Powerlifting*

Sebbene si possano ritrovare delle origini antiche di questo sport, fin dai tempi dell'antica Grecia, il *Powerlifting* moderno è stato riconosciuto solamente negli anni Cinquanta negli Stati Uniti e nel Regno Unito. In quel periodo, infatti, negli Stati Uniti, il *Weightlifting* olimpico perse di interesse, a differenza del *Powerlifting* che iniziò invece a diffondersi. Le alzate che caratterizzano il *Powerlifting* esistono però da molto prima e facevano parte del cosiddetto "*Odd Lifts*", uno sport che coinvolgeva molteplici esercizi di forza.

Durante questi anni di sviluppo, negli Stati Uniti il *Powerlifting* si è diffuso attraverso diversi movimenti. Vi fu, infatti, una chiara distinzione tra i sostenitori del *Weightlifting* olimpico, influenzati da Robert Collins Hoffman, fondatore della rivista *Strength & Health* e proprietario della *York Barbell Company*, allora specializzata nella produzione di attrezzature olimpiche, e i promotori del *Powerlifting* e del *Bodybuilding*, capitanati dall'organizzazione di Joe Weider, co-fondatore dell'*International Federation of BodyBuilders* (IFBB) e creatore delle note competizioni *Mr.* e *Ms. Olympia*. Hoffman, quindi, inizialmente

avversario dell'emergente sport, per contrastare l'influenza di Weider, fondò un'altra rivista denominata *Muscular Development*, maggiormente concentrata sul *Bodybuilding* e sulle competizioni di "*Odd Lifts*". Tra il 1950 e il 1960 diversi eventi "*Odd Lifts*" si spostarono verso le specifiche alzate di distensioni su panca, *squat* e stacco da terra. Hoffman divenne sempre più influente nella diffusione di questo sport tanto che nel 1964 organizzò la prima competizione nazionale negli Stati Uniti, la *Weightlifting Tournament of America*.

Nel frattempo, nel Regno Unito, poiché i membri della *British Amateur Weight Lifters' Association* (BAWLA) erano interessati solo al sollevamento olimpico e non alle nuove alzate emergenti, venne creata una nuova organizzazione, chiamata *Society of Amateur Weightlifters*. All'epoca, i sollevamenti riconosciuti erano 42, ma il cosiddetto "*Strength Set*", composto da *curl* bicipiti, distensioni su panca e *squat*, divenne rapidamente lo standard adottato da entrambe le società nelle competizioni. Nel 1966, per allinearsi con i sollevamenti praticati negli Stati Uniti, il *curl* bicipiti venne sostituito dallo stacco da terra, e la *Society of Amateur Weightlifters* si riunì alla BAWLA. Nello stesso anno si tenne la prima competizione nazionale nel Regno Unito.

Tra la fine degli anni Sessanta e l'inizio degli anni Settanta, si svolsero diverse competizioni amichevoli a livello internazionale e nel 1971 ebbero luogo i primi "*World Weightlifting Championships*".

Nel 1972 si tenne il campionato mondiale dell'*Amateur Athletic Union* (AAU), che vide principalmente la partecipazione di atleti americani. Sebbene l'ordine delle alzate (distensioni su panca, *squat*, stacco da terra) fosse rimasto quello adottato dagli Stati Uniti, non mancò la disapprovazione da parte degli atleti europei. Al termine di questa competizione, i delegati dei vari Paesi si incontrarono per discutere la creazione di un'organizzazione globale per il *Powerlifting*. Lo stesso anno nacque così l'*International Powerlifting Federation* (IPF). Nel 1973, sempre in America, si svolse il primo campionato mondiale organizzato dall'IPF.

La prima gara al di fuori degli Stati Uniti ebbe luogo l'anno seguente a *Birmingham*, in Inghilterra. L'evento ebbe un gran successo, tanto che portò anche a definire delle regole più precise per le future competizioni.

Con i campionati mondiali del 1976 a *York* e quelli del 1977 a *Perth*, si riuscì finalmente a coinvolgere un maggior numero di atleti non provenienti dall'America. Il *Powerlifting* continuò a diffondersi anche negli anni '80 e nel 1982 vennero introdotti, durante i campionati mondiali in Germania, i primi *test antidoping*, che seguirono i principi e requisiti del CIO e che avrebbero avuto lungo in tutte le future competizioni.

Negli stessi anni, iniziò a crescere l'interesse verso questo sport anche da parte delle donne e si svolse così, nel 1980, la prima competizione mondiale femminile. Poco dopo l'IPF divenne membro fondatore della *World Games Association*, l'organizzazione che si occupa degli sport non olimpici e *Powerlifting* venne quindi incluso per la prima volta ai *World Games* nel 1981 a *Santa Clara*, negli Stati Uniti.

Mentre i mondiali maschili e femminili continuarono a svolgersi in tutto il mondo, una novità si registrò nel 1992 quando si tenne per la prima volta un campionato che vedeva coinvolta la singola alzata delle distensioni su panca.

Oggi, dopo decenni di sviluppo e di competizioni, il *Powerlifting* viene praticato in tutto il mondo e rappresenta una disciplina sportiva che continua ad ispirare moltissimi atleti.

A.2 Il *dataset*

A.2.1 Variabili secondarie

Nella sottosezione 2.1.3 sono state presentate le variabili più rilevanti del *dataset OpenPowerlifting*. Si procede ora alla descrizione delle restanti variabili per fornire una visione completa dell'insieme di dati.

- `AgeClass`: la classe di età nella quale ricade l'atleta. Queste classi si basano sull'età esatta dell'atleta il giorno della competizione.
- `BirthYearClass`: la classe di anno di nascita nella quale ricade l'atleta. Solitamente utilizzata dall'IPF e dalle sue federazioni affiliate.
- `WeightClassKg`: la classe di peso in cui l'atleta compete, arrotondato a due decimali. È indicato o come peso massimo (quando è presente solo il numero) o come minimo (indicato con un + a destra del numero).
- `Place`: il posizionamento registrato dell'atleta nella divisione specificata.
  - Numero positivo: il posto raggiunto dall'atleta;
  - `G`: atleta ospite. L'atleta ha avuto successo, ma non era idoneo per i premi;
  - `DQ` e `DD`: rispettivamente squalifica per tentativi falliti o motivi procedurali e per *doping*;
  - `NS`: assente. L'atleta non si è presentato il giorno della competizione.
- `Squat1Kg`, `Squat2Kg`, `Squat3Kg`, `Bench1Kg`, `Bench2Kg`, `Bench3Kg`, `Deadlift1Kg`, `Deadlift2Kg`, `Deadlift3Kg`: rispettivamente il primo, secondo e terzo tentativo di *Squat*, *Bench* e *Deadlift*. I valori negativi sono relativi ai tentativi falliti.

- `Squat4Kg`, `Bench4Kg`, `Deadlift4Kg`: il quarto tentativo di *Squat*, *Bench* e *Deadlift* rispettivamente. I valori negativi sono relativi ai tentativi falliti. I quarti tentativi non contano verso il `TotalKg`. Infatti, solitamente vengono utilizzati per registrare *record* di singoli sollevamenti.
- `Wilks`: un numero positivo se i punti *Wilks* possono essere calcolati, vuoto se l'atleta è stato squalificato.
- `Glossbrenner`: un numero positivo se i punti *Glossbrenner* possono essere calcolati, vuoto se l'atleta è stato squalificato.
- `Goodlift`: un numero positivo se i punti IPF GL possono essere calcolati, è vuoto se l'atleta è stato squalificato o i punti IPF GL non erano definiti per il tipo di evento.
- `Tested`: presenta la modalità "*Yes*" se l'atleta ha partecipato a una categoria sottoposta a *test antidoping* ed è vuota altrimenti. Questa variabile registra se i risultati sono considerati soggetti a *test antidoping*. Questo significa che un atleta potrebbe essere iscritto in una competizione dove possono essere applicati *test antidoping*, ma non è garantito che sia stato testato.
- `Country`: il Paese di origine dell'atleta, se conosciuto.
- `State`: Stato/Provincia/*Oblast*/Divisione di origine dell'atleta, se conosciuto.
- `Federation`: la federazione che ha ospitato la competizione. Questa potrebbe essere diversa dalla federazione internazionale che ha fornito la sanzione per la competizione.
- `ParentFederation`: la federazione principale che ha autorizzato la competizione, di solito l'organismo internazionale.
- `MeetCountry`: il Paese in cui si è svolta la competizione.

- `Sanctioned`: se la competizione conta come ufficialmente autorizzata da una federazione riconosciuta da *OpenPowerlifting*. Le competizioni autorizzate, indicate con la modalità "*Yes*" sono riconosciute da un ente normativo riconosciuto, mentre le competizioni non autorizzate, specificate da "*No*", sono organizzate da direttori di gara individuali, tipicamente al di fuori di una federazione. In *OpenPowerlifting*, le competizioni non autorizzate vengono tracciate, ma non contano per classifiche o *record*.

A.2.2 Punteggi di *performance*

Il *dataset* utilizzato per l'analisi presenta una serie di punteggi per confrontare le prestazioni degli atleti di cui, di seguito, si fornisce la descrizione. Come specificato nella sottosezione 1.2.1, le definizioni sono tratte da OpenPowerlifting Data Service (2024).

I punti DOTS (*Dynamic Objective Team Scoring*) sono molto simili alla formula originale di Wilks. Utilizzano un polinomio aggiornato e più semplice ed è basato sui dati degli atleti *raw* sottoposti a *test antidoping*, invece che su dati di atleti *single-ply* sottoposti a *test antidoping*. La formula è stata creata da Tim Konertz dell'affiliata tedesca della IPF (BVDK) nel 2019.

I punti DOTS permettono di confrontare le prestazioni di atleti, indipendentemente dal sesso e dal peso corporeo. La formula è la seguente:

$$\text{DOTS} = \frac{500}{AW^4 + BW^3 + CW^2 + DW + E} \times T,$$

dove T è il totale dei pesi sollevati dall'atleta in kg, W è il peso corporeo dell'atleta in kg e A, B, C, D, E sono coefficienti specifici definiti in base al sesso dell'atleta (Tabella 1.1).

	Uomini	Donne
A	-0.0000010930	-0.0000010706
B	0.0007391293	0.0005158568
C	-0.1918759221	-0.1126655495
D	24.0900756	13.6175032
E	-307.75076	-57.96288

Tabella A.1: Coefficienti del punteggio $DOTS$.

Il punteggio *Wilks* è basato sul rapporto tra il totale dell'atleta e il peso corporeo dell'atleta in kg. La formula è la seguente:

$$W = \frac{T}{A + B \cdot W + C \cdot W^2 + D \cdot W^3 + E \cdot W^4},$$

dove T è il totale dei pesi sollevati dall'atleta in kg, W è il peso corporeo dell'atleta in kg e A, B, C, D, E sono coefficienti specifici definiti in base al sesso dell'atleta (Tabella A.2).

	Uomini	Donne
A	500	500
B	47.2003	43.5893
C	-0.21004	-0.13743
D	0.00115	0.00034
E	-0.00000355	-0.000000847

Tabella A.2: Coefficienti del punteggio *Wilks*.

Il punteggio *Wilks* viene utilizzato per determinare il miglior atleta nella divisione.

Il *Glossbrenner* è stato creato da Herb Glossbrenner come un aggiornamento della formula di Wilks, in particolare consiste nella combinazione della formula di Schwartz-Malone e il punteggio *Wilks*. È più comunemente usato dalle federazioni affiliate alla GPC. Anche questo punteggio viene utilizzato per il confronto tra *performance* di atleti con diversi pesi corporei.

Per gli uomini:

- quando il peso è inferiore a 153.05kg viene calcolato come

$$\textit{Glossbrenner} = \frac{\text{Schwartz-Malone}(W) + \textit{Wilks-men}(W)}{2}$$

dove Schwartz-Malone(W) è il punteggio di Schwartz-Malone dell'atleta con peso corporeo $W < 153.05$kg e *Wilks-men*(W) è il punteggio *Wilks* calcolato per un atleta maschio con peso $W < 153.05$kg.

- quando il peso corporeo è pari o superiore a 153.05kg viene calcolato come

$$\textit{Glossbrenner} = \frac{\text{Schwartz-Malone}(W) + A \cdot W + B}{2}$$

dove Schwartz-Malone(W) è il punteggio di Schwartz-Malone dell'atleta con peso corporeo $W \geq 153.05$kg, $A = -0.00821668402557$ e $B = 0.676940740094416$.

Per le donne:

- quando il peso è inferiore a 106.3kg viene calcolato come

$$\textit{Glossbrenner} = \frac{\text{Schwartz-Malone}(W) + \textit{Wilks-women}(W)}{2}$$

dove Schwartz-Malone(W) è il punteggio di Schwartz-Malone dell'atleta con peso corporeo $W < 106.3$kg e *Wilks-women*(W) è il punteggio di *Wilks* calcolato per un'atleta femmina con peso $W < 106.3$kg.

- quando il peso corporeo è pari o superiore a 106.3kg viene calcolato come

$$\textit{Glossbrenner} = \frac{\text{Schwartz-Malone}(W) + A \cdot W + B}{2}$$

dove Schwartz-Malone(W) è il punteggio di Schwartz-Malone dell'atleta con peso corporeo $W \geq 106.3$kg, i coefficienti sono pari a $A = -0.000313738002024$ e $B = 0.852664892884785$.

Il punteggio *Goodlift* o IPF GL è il successore dei punti IPF (dal 01-01-2019 al 30-04-2020). I punti IPF GL esprimono, approssimativamente, le prestazioni relative rispetto a quelle attese di quella categoria di peso in un evento di Campionati del Mondo IPF, come percentuale. Si tratta quindi di un metodo di valutazione utilizzato per confrontare le prestazioni di atleti di diverse categorie. La formula risulta:

$$\text{IPF GL} = \frac{100}{A - B \cdot e^{-C \cdot W}} \times T$$

dove W è il peso corporeo dell'atleta, T è il totale sollevato e A, B, C sono coefficienti specifici per sesso, tipo di equipaggiamento ed evento (Tabella A.3).

	Evento	Equipaggiamento	A	B	C
Uomini	SBD	*Raw*	1199.72839	1025.18162	0.009210
		Single-ply	1236.25115	1449.21864	0.01644
	B	*Raw*	320.98041	281.40258	0.01008
		Single-ply	381.22073	733.79378	0.02398
Donne	SBD	*Raw*	610.32796	1045.59282	0.03048
		Single-ply	758.63878	949.31382	0.02435
	B	*Raw*	142.40398	442.52671	0.04724
		Single-ply	221.82209	357.00377	0.02937

Tabella A.3: Coefficienti del punteggio *Goodlift*.

A.2.3 Analisi preliminari

Prima dell'analisi esplorativa, è stato necessario eseguire diverse operazioni di pulizia del *dataset*, principalmente per gestire la presenza di dati mancanti. Infatti, eccetto in due casi, tutte le righe dell'insieme di dati presentano almeno un dato mancante. Inoltre, è stata effettuata una selezione delle variabili, tenendo conto che alcune rappresentano semplicemente una categorizzazione di altre.

Delle tre variabili relative all'età dell'atleta viene mantenuta solamente la terza (`BirthYearClass`), relativa alla categoria di età utilizzata dall'IPF. Per questa variabile, i valori mancanti vengono indicati con la modalità "*missing*". Viene, inoltre, sostituita la modalità "70-999" con "70-99", considerando che la categoria *master* IV è relativa a questa fascia di età.

Date le 93 categorie che la caratterizzavano, è stato poi necessario effettuare un accorpamento della variabile `Division` in modo da ricondursi a quelle riconosciute dall'IPF. Nella maggior parte dei casi, in base al nome delle modalità, è facilmente possibile attribuire le categorie ufficiali. Ad esempio, le modalità `F-O` e `M-O`, come si può intuire, si riferiscono alla categoria *open* di donne e uomini. Ci si è serviti della variabile `BirthYearClass` per verificare la corrispondenza di queste categorie di età.

In altri casi, il nome non è sufficiente per capire la categoria di appartenenza. Per questi, l'accorpamento è stato effettuato utilizzando `BirthYearClass`. Per le unità per le quali `Division` è mancante ma `BirthYearClass` è presente è possibile sostituire i valori assenti con la fascia di età indicata dalla seconda variabile.

Si passa poi alla considerazione delle variabili relative alle alzate svolte dagli atleti. In primo luogo, sono stati rimossi i quarti tentativi, non incidenti nella classifica. Nelle prime tre alzate compaiono molti valori mancanti, che di fatto non corrispondono a dati realmente assenti. Infatti, in alcune competizioni non vengono effettuate tutte e

tre le alzate. Si decide, dunque, di sostituire i valori mancanti con 0. Inoltre, per gli atleti con almeno una prova fallita, indicata con segno negativo, non viene segnalato il miglior tentativo. Per questi casi, è possibile sostituire i valori mancanti delle tre variabili `Best3SquatKg`, `Best3BenchKg`, `Best3DeadliftKg` con il massimo osservato nelle tre prove delle alzate prese singolarmente. È da specificare che se tutti i valori nelle singole alzate sono negativi, ossia se ci sono solamente tentativi falliti, allora la miglior alzata dovrà essere nulla e, di fatto, l'atleta è squalificato (`Place` pari a `DQ`).

A seguito di queste sostituzioni, è possibile ricalcolare il valore della variabile `TotalKg` come somma di `Best3SquatKg`, `Best3BenchKg`, `Best3DeadliftKg`, in modo da non avere valori mancanti. In alcuni casi, però, non sono presenti dati relativi alle singole alzate ma solamente il totale. Per questi atleti, procedendo al ricalcolo, il vero valore di `TotalKg` verrebbe sostituito (sbagliando) da zero. Si dovranno, quindi, escludere questi atleti e ricalcolare il totale sollevato solamente per i restanti. Nuovamente, si dovrà tener conto che se un atleta ha solamente tentativi falliti allora il totale sollevato sarà 0.

A partire dai valori delle alzate è possibile creare una nuova variabile relativa al numero di tentativi falliti.

Focalizzandosi sulla variabile `Place`, vengono mantenute le prime cinque posizioni, accorpando le seguenti in una nuova modalità "5+". Si decide di eliminare gli atleti assenti dalle competizioni (`Place` pari a `NS`), dato che questi non sono di interesse per l'analisi.

Come descritto in 1.2.1 e A.2.2, sono a disposizione diversi indicatori per il confronto delle *performance* degli atleti. Si decide di considerare solamente `Dots`, solitamente utilizzato e visualizzato durante le competizioni. Per questo punteggio si hanno molti valori mancanti, in corrispondenza degli atleti squalificati e di quelli per i quali non è registrato il peso. Innanzitutto si procede a sostituire il valore mancante di atleti squalificati con 0 mentre per gli atleti in cui non è stato indicato il peso si segue questa strategia:

- si sostituisce il dato mancante con la mediana dei valori osservati in corrispondenza della medesima categoria di peso (qualora presenti). Ad esempio, se per l'atleta `X` non è presente un valore in corrispondenza di una categoria di peso 140+, si prenderà la mediana dei pesi registrati per le altre competizioni che ha svolto nella categoria 140+;
- se non sono presenti valori di peso in corrispondenza della categoria per la quale si ha il dato mancante, si valuta se esiste una categoria di peso superiore. Se presente, si sostituirà il valore mancante con la mediana dei pesi appartenenti a categorie superiori, che risulteranno comunque vicine alla vera categoria di peso. Ad esempio, se l'atleta `X` presenta un dato mancante in corrispondenza di una categoria di peso 140+ ma non sono presenti altri valori appartenenti a quella categoria, si valuta se ce ne sono per una superiore, come 145, e si utilizza la mediana dei pesi ad essa relativi;
- se non sono presenti nemmeno valori contenuti in categorie superiori, si sostituisce il valore mancante con il peso che rappresenta il limite inferiore o superiore della classe di peso nella quale l'atleta compete. Ad esempio, se il soggetto `X` presenta un valore mancante in corrispondenza di una categoria di peso 140+ ma non sono presenti valori di peso in altre competizioni per quella o altre categorie superiori, si inputa con 140.

La procedura di sostituzione dei valori mancanti si basa sull'ipotesi che sia raro che un atleta cambi radicalmente la categoria di peso da una competizione all'altra. Si sottolinea, inoltre, che l'adozione della mediana è stata effettuata per assicurare una maggiore robustezza.

Il peso è mancante anche per tre atleti squalificati. Si procede in modo analogo a quanto appena descritto anche se il punteggio `Dots` non verrà calcolato. Inoltre, in 11 casi non è stato registrato né peso né categoria di peso. Questi atleti sono stati esclusi dell'analisi.

Terminata la gestione di dati mancanti per il peso è stato possibile procedere al calcolo dei punti `Dots`.

La prossima variabile da esaminare è `WeightClassKg`, la categoria di peso nella quale l'atleta compete. Anche per questa, così come per `BirthYearClass`, vi sono molte modalità (58). Vengono mantenute solamente quelle riconosciute dall'IPF, riportate nella Tabella 1.1. Per effettuare la giusta suddivisione viene utilizzata l'informazione fornita dalla variabile `BodyweightKg`. Si ricorda che, in alcuni casi, specificatamente per 29 atleti, il peso è stato inputato per poter calcolare i punti `Dots`.

Anche la variabile `Tested` presenta dati mancanti in corrispondenza delle categorie che vengono sottoposte a *test antidoping*. È quindi sufficiente sostituire questi valori con la modalità "*No*".

Le due variabili successive sono `Federation` e `ParentFederation`, che indicano, rispettivamente la federazione che ha ospitato la competizione e l'ente che l'ha autorizzata. Si mantiene solo la seconda variabile, i cui valori mancanti sono inglobati nella modalità "*missing*".

Tra le informazioni relative agli atleti sono presenti il Paese e lo Stato di origine. Data la varietà di modalità del primo, è stato effettuato un accorpamento mantenendo solo quelle più comuni, ossia con una frequenza superiore a 200, passando da 66 a 9 modalità.

La variabile `Country`, che rappresenta lo Stato di origine dell'atleta, è stata eliminata poiché contiene esclusivamente valori mancanti.

Le ultime informazioni contenute nel *dataset* riguardano le competizioni. Dalla variabile relativa alla data delle gare viene estratto l'anno. Si è deciso di eliminare le variabili `MeetCountry`, in quanto, avendo selezionato le competizioni italiane, presenta solo una modalità, e `MeetState`, che contiene esclusivamente valori mancanti.

Per quanto riguarda la variabile `MeetTown`, è stato necessario uniformare la scrittura di alcune città ed è stato corretto qualche errore di battitura. La variabile è stata poi sostituita da latitudine e longitudine, allo scopo di fornire un riferimento geografico più specifico.

Nel caso in cui non sia stata indicata la città in cui si è tenuta la competizione, anche le due nuove variabili presenteranno valori mancanti, che verranno quindi classificati come "*missing*".

Infine, è stata eliminata la variabile `Sanctioned`, poiché presenta unicamente la modalità "*Yes*", indicando che tutte le competizioni sono riconosciute da una federazione accreditata da *OpenPowerlifting* e, pertanto, non offre informazioni utili.

La numerosità dell'insieme di dati passa così da 30651 a 30615 mentre la dimensionalità si riduce a 30 variabili.

Appendice B

Materiale aggiuntivo capitolo 2

B.1 Statistiche descrittive

In questa sezione verranno forniti alcuni approfondimenti sulle statistiche descrittive esaminate in 2.1.2.

Il primo indicatore analizzato è il grado, di cui è stata riportata la formula per reti dirette e indirette. Quando non vi è direzionalità nella connessione, l'arco presenta due estremità. Questo significa che se ci sono m archi, si avranno $2m$ estremità. Questo numero risulta anche pari alla somma dei gradi di tutti i nodi presenti nella rete, vale a dire

$$2m = \sum_{i=1}^{n} k_i = \sum_{i,j} A_{ij}.$$

Nelle reti dirette, invece, è possibile calcolare il grado in entrata e quello in uscita. Se il numero di archi in una rete diretta è m si ha che

$$m = \sum_{i=1}^{n} A_{ij} = \sum_{j=1}^{n} A_{ij}.$$

Ne consegue che la media del grado in entrata e di quello in uscita coincidono.

Tra le misure per valutare la posizione di un nodo all'interno di una rete vi è la *closeness centrality* che, in generale, si può ottenere tramite 2.1. Ciò nonostante, nel caso di reti con più componenti, considerato che la distanza tra due nodi appartenenti a diversi sotto-grafi è infinita, C_i risulterebbe nulla. Una soluzione semplice consiste nel calcolare la media solo per nodi all'interno di una stessa componente. Tuttavia, in questo modo, componenti di dimensioni ridotte porterebbero a valori piccoli di ℓ_i e, quindi, ad un'elevata *closeness centrality* rispetto a nodi di componenti più grandi. Questo conduce all'utilizzo della media armonica delle distanze, piuttosto che quella aritmetica, ridefinendo C_i come

$$C_i' = \frac{1}{n-1} \sum_{j(\neq i)} \frac{1}{d_{ij}}.$$

L'esclusione di d_{ii} dalla somma, a differenza di 2.1, è necessaria per non far sì che $C_i = \infty$. In questo modo, per due nodi appartenenti a componenti differenti, ovvero con $d_{ij} = \infty$, il reciproco della distanza sarà nullo, escludendo la coppia di nodi dalla somma.

Un'altra misura di centralità considerata è la *betweenness centrality*. La formula in 2.2 può essere utilizzata sia per reti dirette che per quelle indirette. Si specifica però che, poiché gli archi in una rete indiretta non hanno direzionalità, i percorsi vengono contati due volte. Ad ogni modo, l'indicatore non viene normalizzato per poter utilizzare la stessa formula anche nel caso di reti dirette. Questo non influisce sulla sua interpretazione, dato che ciò che conta è principalmente la grandezza relativa della centralità piuttosto che il suo valore assoluto (Newman, 2018).

B.1.1 Proprietà

Una delle proprietà fondamentali delle reti è il fenomeno del mondo piccolo, che suggerisce che, nonostante le dimensioni considerevoli comunemente riscontrate nelle reti reali, la distanza media tra i nodi è inaspettatamente corta.

Approfondendo maggiormente si può capire che l'effetto del mondo piccolo in realtà non è così inatteso. Infatti, i modelli matematici delle reti casuali suggeriscono che le lunghezze dei percorsi osservati in una rete si comportano, all'incirca come $\log n$, dove n è il numero di nodi. Considerato che il logaritmo è una funzione che cresce lentamente, questo fa sì che all'aumentare del numero di nodi la lunghezza dei percorsi tra coppie di essi rimanga contenuta (Newman, 2018).

Nel caso di reti dirette, il percorso che parte dal nodo i e arriva al nodo j è, generalmente, diverso da quello in direzione opposta. Pertanto, i percorsi che collegano i e j potrebbero avere lunghezze diverse, oppure potrebbe non esistere un percorso in una o in entrambe le direzioni. Come nel caso di reti indirette con più componenti il problema può essere risolto sfruttando la formula in 2.3. Questa non può invece essere utilizzata nel caso di reti con diverse componenti. Il cammino che collega una coppia di nodi appartenenti a sotto-grafi differenti non esiste ($d_{ij} = \infty$), il che rende impossibile il calcolo di ℓ tramite 2.3.

In queste circostanze, quindi, la media delle distanze più brevi viene sostituita da

$$\ell = \frac{\sum_m \sum_{i,j \in \mathcal{C}_m} d_{ij}}{\sum_m n_m^2},$$

dove n_m è il numero di nodi nella componente m, indicata con $\mathcal{C}_m$.

Un altro elemento chiave per l'analisi delle reti è la distribuzione del grado che fornisce informazioni sulla connettività tra i nodi. La distribuzione del grado viene solitamente utilizzata per realizzare un grafico, come quello in Figura B.1, che pone la frazione p_k di nodi con grado k in funzione di quest'ultimo. La maggior parte delle reti mostra una distribuzione con una lunga coda a destra, ad indicare la presenza di pochi nodi con molte connessioni. Questi prendono il nome di *hub*.

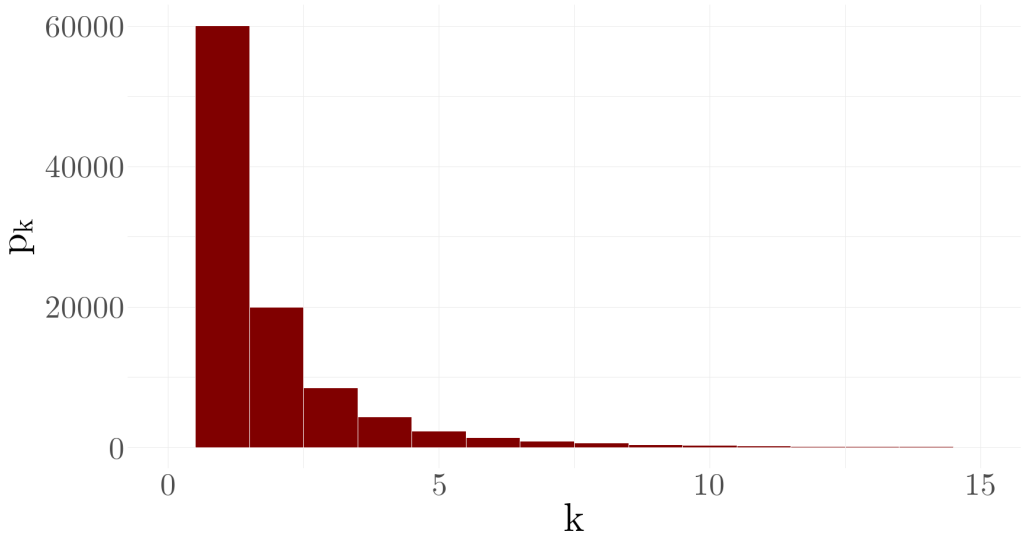


Figura B.1: Distribuzione del grado in una rete indiretta simulata con $n = 100000$ nodi.

È possibile calcolare la distribuzione del grado anche nel caso di reti dirette. In questo caso, si parla di grado in entrata e in uscita per indicare, rispettivamente, il numero di archi che arrivano e quelli che partono da un determinato nodo. Si potranno dunque realizzare due grafici distinti. Tuttavia, per essere più specifici, nel caso di reti dirette la distribuzione del grado dovrebbe essere costituita dalla densità congiunta del grado in entrata e di quello in uscita. Si indica con p_{ij} la frazione di nodi che presenta un grado in entrata pari a i e un grado in uscita di j. In questo modo, si terrà conto della correlazione presente tra le due tipologie di grado, che non verrebbe considerata analizzando le due distribuzioni singolarmente.

Utilizzando la scala logaritmica in entrambi gli assi del grafico della distribuzione del grado si può notare un andamento lineare con pendenza negativa. Ne consegue che il logaritmo della distribuzione del grado p_k è funzione lineare del logaritmo del grado k in modo tale che

$$\ln p_k = -\alpha \ln k + c, \tag{B.1}$$

dove α e c sono costanti e il segno meno, anche se opzionale, rappresenta la pendenza negativa, il che implica un valore positivo per α.

Applicando la funzione esponenziale in entrambi i membri di B.1 si ottiene

$$p_k = C \cdot k^{-\alpha}, \tag{B.2}$$

dove C è una costante di normalizzazione. Molte distribuzioni seguono questa forma, che varia come potenze di k. Queste sono denominate leggi di potenza (*power laws*) di cui α, che solitamente assume valori compresi tra 2 e 3, è l'esponente.

In generale, le reti potrebbero mostrare una distribuzione del grado diversa da B.2 lungo l'intero intervallo, mostrando, ad esempio, la riduzione di p_k per valori piccoli di k. In senso stretto, la vera distribuzione di potenza risulta, invece, monotona decrescente, per ogni valore di k. È usuale che questa caratteristica venga rispettata solamente per valori elevati del grado, ma si fa comunque riferimento a tali distribuzioni come leggi di potenza.

Le reti che presentano questa tipologia di distribuzioni del grado vengono anche chiamate reti *scale-free*, anche se una rete con legge di potenza potrebbe non essere *scale-free*. Identificare questo tipo di reti non è semplice e spesso si utilizza il grafico che vede $\log p_k$ in funzione di $\log k$ per identificarle. Il problema principale di questo approccio è la limitata presenza di nodi valori elevati del logaritmo del grado k, che provoca rumore nella coda destra della distribuzione, rendendo difficile comprendere se quest'ultima segua o meno una retta.

Una prima soluzione è quella di utilizzare una larghezza delle barre più ampia in modo da includere un maggior numero di nodi. Questo riduce il rumore ma porta ad una minore precisione. Una seconda alternativa è invece quella di utilizzare larghezze differenti al crescere dei valori sull'asse delle ascisse: aumentando la larghezza delle barre nella parte sinistra della distribuzione e diminuendo quella sulla destra, si riuscirebbe a ridurre il rumore mantenendo al tempo stesso una certa precisione.

Una terza alternativa è data dalla funzione di distribuzione cumulativa del grado

$$P_k = \sum_{k_0=k}^{\infty} p_{k_0},$$

dove P_k indica la frazione di nodi con grado maggiore o uguale a k. La distribuzione cumulativa in scala logaritmica verrà rappresentata in funzione del logaritmo del grado, così come in Figura B.2, da cui è possibile verificare la presenza di un andamento lineare.

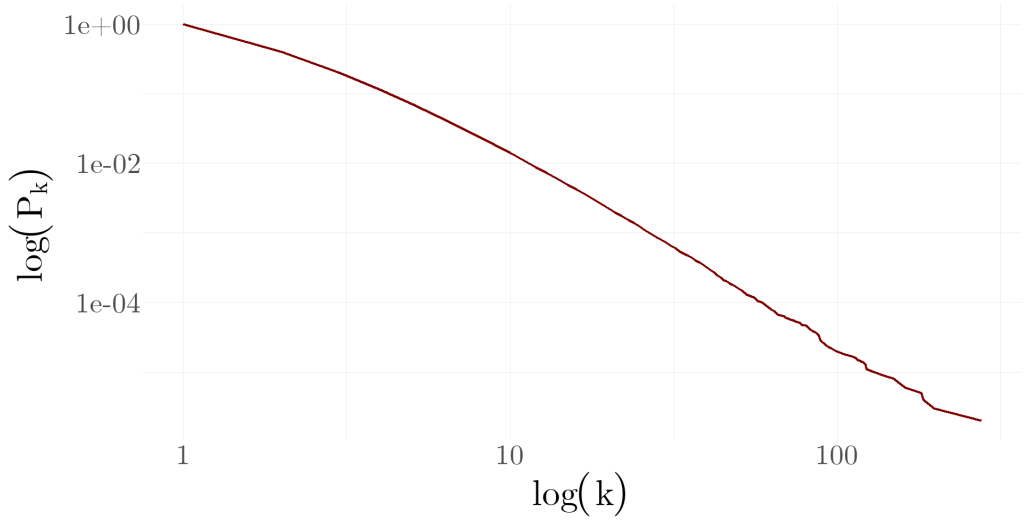


Figura B.2: Distribuzione cumulativa in funzione del grado (in scala logaritmica) in una rete indiretta simulata con $n = 100000$ nodi.

L'utilizzo di P_k presenta diversi vantaggi. Oltre alla semplicità di calcolo, l'assenza di barre consente di mantenere tutta l'informazione presente nei dati, eliminando la necessità di determinare come suddividere i valori di k. Sono presenti, tuttavia, alcuni svantaggi: una maggiore difficoltà di interpretazione e la dipendenza tra punti successivi del grafico.

Un'altra importante proprietà è la transitività, che riflette la tendenza dei nodi a formare dei gruppi tramite le loro connessioni. Questo concetto viene misurato mediante il coefficiente di *clustering* C.

Si supponga che il nodo i sia connesso a j e j presenti un arco che lo congiunge a m. Questo significa che vi è un percorso che parte da i, passa per j ed arriva a k, con una lunghezza pari a 2. Nel caso in

cui anche i e k siano connessi, il percorso, che avrà ora lunghezza tre, sarà chiuso e formerà un triangolo nella rete. In questa circostanza, i, j e k costituiscono una triade chiusa. È possibile, quindi, definire il coefficiente di *clustering* come il numero di percorsi chiusi di lunghezza due sul totale dei percorsi con questa lunghezza, ossia

$$C = \frac{\text{numero di percorsi chiusi di lunghezza due}}{\text{numero di percorsi di lunghezza due}}.$$

Il coefficiente di *clustering* sarà pari a 1 nel caso di transitività perfetta e 0 in assenza di triadi chiuse.

Una formulazione alternativa è la seguente

$$C = \frac{\text{numero di triangoli} \cdot 6}{\text{numero di percorsi di lunghezza due}},$$

dove 6 è inserito considerando che ogni triangolo conterrà sei percorsi di lunghezza due.

La formulazione più utilizzata per C è, tuttavia, la seguente:

$$C = \frac{\text{numero di triangoli} \cdot 3}{\text{numero di triplette connesse}},$$

in cui 3 si riferisce al modo con cui i triangoli vengono contati all'interno di una tripletta connessa, dove quest'ultima consiste in un insieme di tre nodi, i, j e k con archi $\{i, j\}$ e $\{j, k\}$.

Questa formulazione è dovuta al fatto che, se è presente un percorso di lunghezza due che congiunge i, j e k, allora i e k avranno un vicino comune (j). Se il percorso tra questi tre nodi fosse chiuso, allora anche i e k sarebbero collegati. Il coefficiente di *clustering* può, dunque, essere inteso come la probabilità media che due nodi con un vicino in comune siano connessi. In generale, le reti sociali mostrano valori elevati del coefficiente di *clustering*.

B.2 Reti bipartite

Come per le reti unipartite, è possibile calcolare alcune statistiche descrittive anche per la struttura di rete bipartita. Considerando una rete generica che coinvolge persone ed eventi, in questa sezione ci si riferirà alle prime come attori, al fine di rendere più chiara la distinzione tra le due tipologie di nodi.

Il primo indicatore visto nella sottosezione 2.1.2 è il grado. Anche nel caso di reti bipartite è possibile procedere al calcolo di questa misura. Tuttavia, poiché le reti bipartite coinvolgono due categorie distinte di nodi, è opportuno esaminare ciascuna di esse separatamente. Un nodo sarà considerato centrale in base al numero di connessioni che possiede e alla frequenza con cui partecipa agli eventi, mentre gli eventi stessi saranno considerati rilevanti in funzione del numero di partecipanti che vi prendono parte.

In una rete bipartita, la centralità di grado può essere calcolata in diversi modi, a seconda che si utilizzi la matrice di incidenza o le sue proiezioni monomodali. Focalizzandosi nel primo caso, il grado dei nodi attori è dato dalla somma dei valori sulla diagonale della matrice di co-affiliazione, mentre per gli eventi corrisponde alla somma degli stessi elementi nella matrice di sovrapposizione degli eventi. Se si considera la matrice di affiliazione, questi valori coincidono rispettivamente con la somma effettuata per riga e per colonna. In altre parole, l'indice di centralità di grado per gli attori rappresenta il numero di eventi a cui hanno partecipato, mentre per gli eventi riflette il numero di attori che vi hanno preso parte.

L'indice di centralità di grado k_i per gli attori e per gli eventi risulta rispettivamente

$$k_i^A = \sum_{j=1}^{g} B_{ij}, \quad k_j^E = \sum_{i=1}^{n} B_{ij},$$

dove B_{ij} è il generico elemento in posizione (i, j) della matrice di incidenza.

Utilizzando le proiezioni monomodali, si dimostra che la centralità di grado per gli attori è data dalla somma del numero degli eventi a cui partecipano, mentre la centralità di grado degli eventi è uguale alla somma delle appartenenze degli attori coinvolti. Questi risultati evidenziano una relazione diretta tra la centralità degli attori e quella degli eventi nella rete bipartita, confermando che il grado di un attore dipende dalla partecipazione a eventi e viceversa (Faust, 1997).

B.3 Reti dinamiche

Vi è una relazione tra rete dinamica e multistrato. Una rete multistrato è composta da un insieme di singole reti, denominate livelli, ciascuna contenente una determinata tipologia di nodi e connessioni. In una rete multistrato possono essere presenti anche archi interstrato, che connettono nodi appartenenti a livelli differenti. Da questa definizione è chiaro che anche le reti bipartite, descritte nella sottosezione 2.2.1, rientrano in questa classificazione. Anche una rete dinamica può essere vista come una rete multistrato in cui ogni livello rappresenta un istante temporale. Se l'insieme di nodi è lo stesso per ogni strato, si parla di rete *multiplex*, se, invece, i nodi sono diversi in ciascun livello, ovvero possono apparire e scomparire, si farà riferimento ad una rete multistrato. In quest'ultimo caso, è necessario aggiungere archi interstrato per indicare l'equivalenza dei nodi (Newman, 2018).

Nel contesto delle reti dinamiche, le misure descritte in 2.1.2 richiedono un adattamento per tenere conto della variabilità temporale delle connessioni. Il grado, ad esempio, può essere calcolato semplicemente considerando le connessioni presenti in un determinato intervallo. Altre statistiche, come la *shortest distance*, devono essere ridefinite in termini di attivazione degli archi, ovvero considerando le sequenze di connessioni con un certo ordine temporale, in modo da rispettare il momento in cui ogni arco si presenta. Questi percorsi prendono il nome di *time-respecting paths* e portano a definire i due concetti di

set di influenza e di origine di un nodo i. Il primo è l'insieme di nodi che possono essere raggiunti da i. Il secondo, al contrario, rappresenta l'insieme di nodi che possono raggiungerlo (Holme, 2012). I percorsi che rispettano il tempo portano anche alla non transitività delle connessioni. In una rete statica, se il nodo i è connesso al nodo j e j è legato a k, allora indirettamente si avrà una connessione tra i e k. Nel caso di reti dinamiche, la situazione è più complessa e la transitività non è garantita a causa del fattore temporale. In questo contesto, le interazioni possono avvenire in momenti diversi, e un collegamento che esiste in un determinato istante potrebbe non essere valido in un altro.

Appendice C

Materiale aggiuntivo capitolo 3

C.1 Algoritmi *force-directed*

Nel Capitolo 3 sono stati introdotti due degli algoritmi di *force-directed placements* più utilizzati.

L'algoritmo Fruchterman-Reingold si basa su un modello fisico di interazione tra i nodi attraverso forze attrattive e repulsive, cercando la configurazione che minimizzi l'energia presente. Indicando le due forze, rispettivamente, con f_r e f_a si ha

$$f_r(i,j) = -\frac{CK^2}{\|x_i - x_j\|}, \quad \forall i \neq j, \quad i,j \in V,$$
$$f_a(i,j) = \frac{\|x_i - x_j\|^2}{K}, \quad \forall i \leftrightarrow j,$$

dove x_i e x_j indicano le coordinate dei nodi i e j, $\|x_i - x_j\|$ è la norma euclidea, $i \leftrightarrow j$ indica che i e j sono connessi, K è il parametro di distanza ottimale e C regola la forza relativa delle forze repulsive e attrattive.

Fruchterman e Reingold hanno arricchito il lavoro originario di Eades tramite l'introduzione del concetto di temperatura, che regola la grandezza degli spostamenti dei nodi durante il processo di ottimiz-

zazione. Man mano che la temperatura diminuisce, la mobilità dei nodi viene progressivamente ridotta, stabilizzando il *layout* mentre si avvicina a una configurazione ottimale (Kobourov, 2013).

Il fatto che l'algoritmo simuli il comportamento fisico di molle e cariche elettriche per il posizionamento dei nodi chiarisce perché venga anche identificato come modello elettrico a molla (*spring-electrical model*). Questi modelli, soprattutto nel caso di reti con un grande diametro, presentano il cosiddetto effetto periferico, in cui i nodi situati ai margini tendono ad essere disposti più vicini tra loro rispetto a quelli al centro. Questo fenomeno è causato dalla forza repulsiva, che agisce anche su nodi molto distanti tra loro. Sebbene l'effetto di questa forza diminuisca con la distanza, il suo decadimento è abbastanza lento da provocare tale effetto. Tuttavia, nella maggior parte dei casi, la configurazione generale non subisce un impatto significativo.

Nel metodo di Kamada e Kawai, invece, l'obiettivo è quello di minimizzare l'energia in termini di differenza tra distanza osservata e teorica. La *shortest distance*, d_{ij}, rappresenta la distanza teorica tra i due nodi, mentre la distanza ideale dell'arco è $l_{ij} = L \times d_{ij}$, dove il suggerimento degli autori è quello di utilizzare un valore di L pari al rapporto tra lunghezza del lato dell'area nella quale viene visualizzata la rete e la massima *shortest distance*, ossia il diametro.

La forza della molla, k_{ij}, sarà dunque

$$k_{ij} = \frac{K}{d_{ij}^2},$$

dove K è una costante. L'energia complessiva del sistema è data dalla somma dei quadrati delle differenze tra le distanze geometriche osservate e quelle teoriche per ogni coppia di nodi, pesate in base alla forza della molla che li collega.

A differenza dell'algoritmo Fruchterman-Reingold, il modello di Kamada-Kawai non genera l'effetto periferico.

C.2 Confronto tra algoritmi

Nella sottosezione 3.3.1 vengono visualizzate le reti di partecipazione competitiva nel *Powerlifting* maschile nel 2023 utilizzando algoritmi *force-directed* per il posizionamento dei nodi. Rispetto alla disposizione casuale, questi metodi si sono dimostrati efficaci per l'interpretazione delle interazioni tra gli atleti. Tuttavia, i due algoritmi utilizzati sono solo alcuni dei metodi disponibili nella libreria `igraph`. In questa sezione ne verranno considerati altri tre allo scopo di giustificare l'utilizzo di Fruchterman-Reingold nelle analisi presentate.

Dalla Tabella C.1, si può vedere come i tempi di esecuzione variano significativamente tra i diversi algoritmi. La configurazione casuale dei nodi risulta chiaramente la più rapida mentre tra gli algoritmi più complessi, Fruchterman-Reingold si distingue come il più efficiente, impiegando meno di un secondo.

Layout	**Tempo di esecuzione**
Random	0.00
Fruchterman-Reingold	0.60
Kamada-Kawai	6.70
Multidimensional Scaling	11.75
Force-directed	80.64
Large Graph	491.78

Tabella C.1: Tempo di esecuzione in secondi dei principali algoritmi disponibili in `igraph`.

Dalle diverse configurazioni in Figura C.1, si può notare come la disposizione casuale, così come i metodi di *Multidimensional Scaling* e di *Force-Directed* non contribuiscano in modo sufficiente nella rappresentazione delle interazioni tra gli atleti. Non si evidenziano infatti atleti con alti valori di *betweenneess* che potrebbero risultare più competitivi per la loro maggiore presenza alle competizioni, gruppi fortemente connessi o ancora comunità più distanti. Le visualizzazioni che chiariscono meglio questi aspetti sono quelle generate dagli

algoritmi Fruchterman-Reingold e *Large Graph*, le cui configurazioni appaiono simili, sebbene il secondo metodo mostri una simmetria più accentuata nella comunità di dimensione maggiore. Alla luce di questo e della velocità computazionale richiesta, l'algoritmo Fruchterman-Reingold si è dimostrato la scelta più adeguata per l'analisi delle reti di partecipazione nel *Powerlifting*.

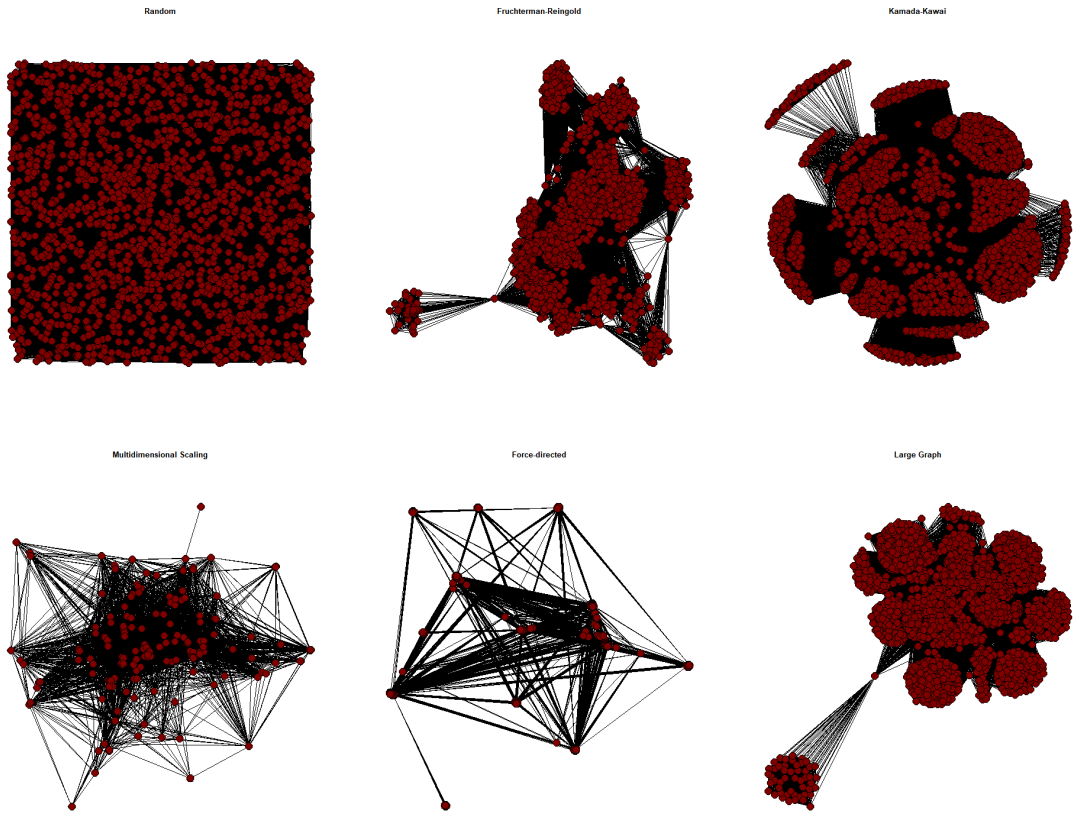


Figura C.1: Reti maschili del 2023: disposizione dei nodi utilizzando i principali algoritmi disponibili in `igraph`.

C.3 *Community detection*

In questa sezione, facendo riferimento a Aynaud and Guillaume (2010), Newman and Girvan (2004) e Newman (2018) e a quando approfondito nella sezione 3.2, verrà riportato lo pseudo-codice degli algoritmi di Louvain e Girvan-Newman.

Algorithm 1 Algoritmo di Louvain

Sia $G = \{V, E\}$.

1. Inizializzazione: ogni nodo $i \in V$ è una singola comunità.
2. Iterazione:
   - **Fase 1**: ripetere fino a quando si verificano spostamenti.
     (a) Calcolare il guadagno di modularità $\Delta Q(i \to j)$ per ogni possibile spostamento del nodo i nella comunità j.
     (b) Spostare il nodo i nella comunità j che massimizza il guadagno di modularità, ossia $j : Q = \max_j \Delta Q(i \to j)$.

   Se la nuova modularità è maggiore della modularità iniziale:
   - **Fase 2**: creare un nuovo grafo G i cui nodi sono le comunità identificate nella Fase 1.

Algorithm 2 Algoritmo di Girvan-Newman

Sia $G = (V, E)$.

1. Calcolare la *betweenness* per tutti gli archi della rete G.
2. Identificare l'arco con il valore di *betweenness* più alto e rimuoverlo dall'insieme E. Se più archi hanno lo stesso valore dell'indicatore, selezionarne uno casualmente.
3. Ricalcolare la *betweenness* per gli archi rimanenti.
4. Ripetere il passo 2 fino a quando non ci sono più archi da rimuovere.

Elenco delle figure

Elenco delle tabelle

Bibliografia

Avrachenkov, K. and Dreveton, M. (2022). *Statistical Analysis of Networks*, *Now Publishers*. https://www.nowpublishers.com/article/BookDetails/9781638280507.

Aynaud, T. and Guillaume, J.-L. (2010). *Static community detection algorithms for evolving networks*. In: 8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks, pp. 513-519. https://ieeexplore.ieee.org/document/5520221.

Breiger, R.L. (1974). *The Duality of Persons and Groups*, *Social Forces*. 53(2), pp. 181-190. https://doi.org/10.2307/2576011.

Csárdi, G., et al. (2024). *Network analysis and visualization*. https://cran.r-project.org/web/packages/igraph/igraph.pdf.

Faust, K. (1997). *Centrality in affiliation networks*, *Social Networks*. 19(2), pp. 157-191. https://doi.org/10.1016/S0378-8733(96)00300-0.

Filho, A.D.V. (2022). *Introduction to historical (social) network analysis – Part I*, *Digital Humanities Lab*. https://doi.org/10.58079/nl66.

Fruchterman, T.M.J. and Reingold, E.M. (1991). *Graph drawing by force-directed placement*. *Software Practice and Experience*. 21(11), pp. 1129-1164. https://doi.org/10.1002/spe.4380211102.

Hoff, P. D. (2015). *Dyadic data analysis with amen.* https://arxiv.org/abs/1506.08237.

Hoff, P.D. (2018). *Additive and multiplicative effects network models.* https://arxiv.org/abs/1807.08038.

Hoff, P., et al. (2024). *Additive and Multiplicative Effects Models for Networks and Relational Data.* https://cran.r-project.org/web/packages/amen/amen.pdf.

Holme, P. and Saramäki, J. (2012). *Temporal networks*, *Physics Reports.* 519(3), pp. 97-125. https://doi.org/10.1016/j.physrep.2012.03.001.

IPF (2023). Technical Rules Book, Technical Rules book of the International Powerlifting Federation. https://www.powerlifting.sport/fileadmin/ipf/data/rules/technical-rules/english/IPF_Technical_Rules_Book_2023__1_.pdf

IPF (2024). *History.* https://www.powerlifting.sport/federation/history

O'g'li, J.S.S. (2022). *The history of the origin and development of the sport of powerlifting. Zenodo (CERN European Organization for Nuclear Research)*, 1(17), pp. 155-158. https://zenodo.org/records/7312473.

Kamada, T. and Kawai, S. (1989). *An algorithm for drawing general undirected graphs. Information Processing Letters*, 31, pp. 7-15. http://itginsight.com/wp-content/uploads/2022/09/AN-ALGORITHM-FOR-DRAWING-GENERAL-UNDIRECTED-GRAPHSKadama-Kawai-layout.pdf.

Kobourov, S. (2013). *Force-Directed Drawing Algorithms.* In Tamassia, R. *Handbook of Graph Drawing and Visualization*, pp. 383-408.

Boca Raton, FL: CRC Press. https://api.semanticscholar.org/CorpusID:13427672.

Kolaczyk, E.D. and Csárdi, G. (2020). *Statistical Analysis of Network Data with R*. https://doi.org/10.1007/978-3-030-44129-6.

Lizardo, O. and Jilbert, I. (2023). *Social networks*. https://olizardo.github.io/networks-textbook/.

Newman, M.E.J. and Girvan, M. (2004). *Finding and evaluating community structure in networks*. *Physical Review E*, 69(2). https://doi.org/10.1103/physreve.69.026113.

Newman, M. (2018). *Networks*. *Oxford University Press*. https://doi.org/10.1093/oso/9780198805090.001.0001.

OpenPowerlifting (2024). *Powerlifting Rankings*. https://www.openpowerlifting.org/ (accessed on April 3, 2024).

OpenPowerlifting Data Service (2024). *Documentation*. https://openpowerlifting.gitlab.io/opl-csv/bulk-csv-docs.html.

PowerliftingToWin (2014a). *Powerlifting bench press rules*. https://www.powerliftingtowin.com/powerlifting-rules-bench-press/.

PowerliftingToWin (2014b). *Powerlifting deadlift rules*. https://www.powerliftingtowin.com/powerlifting-rules-deadlift/.

Sarmento, R. P., Lemos, L., Cordeiro, M., Rossetti, G., & Cardoso, D., 2019. *DynComm R Package - Dynamic community Detection for evolving networks*. https://arxiv.org/abs/1905.01498.

PowerliftingToWin (2014c). *Powerlifting squat rules*. https://www.powerliftingtowin.com/powerlifting-squat-rules/.

Yifan, H., 2006. *Efficient, High-Quality Force-Directed Graph Drawing. The Mathematica journal*, 10, pp. 37-51. https://api.semanticscholar.org/CorpusID:14599587.

Ringraziamenti

Questo elaborato segna la conclusione di un lungo e importante percorso. Vorrei dunque dedicare qualche riga a ringraziare le persone senza le quali tutto questo non sarebbe risultato possibile.

Ringrazio innanzitutto la mia relatrice, Prof.ssa Mariangela Guidolin, per la sua professionalità, disponibilità e la grande gentilezza che ha sempre dimostrato.

La mia più profonda gratitudine è rivolta ai miei genitori, sempre pronti a supportarmi e ad incoraggiarmi in ogni momento. Vi ringrazio per l'enorme affetto dimostrato e per aver accolto il mio carattere spesso complicato. Un grazie va anche a mio fratello Giacomo, nonché (finalmente) collega statistico, per avermi affiancato in parte di questo percorso e per essere sempre pronto ad aiutarmi.

Un ringraziamento particolare va anche a tutti i miei parenti e un pensiero è rivolto a nonno Romano e zia Monica che continuerò a portare nel mio cuore.

Un grazie speciale a Martina che con premura non ha mai dimenticato di farmi sentire il suo supporto per ogni esame.

Ringrazio Camilla, ormai sorella acquisita, e le mie care amiche Eleonora e Beatrice per la costante presenza e per trovare sempre le parole giuste per confortarmi.

Grazie anche ad Alex, per credere in me più di quanto io stessa riesca e per farmi sentire all'altezza.

A Mirko, compagno di questo lungo viaggio, un immenso grazie per essere sempre stato al mio fianco e per riuscire sempre a strapparmi un sorriso. Grazie perché con la tua presenza e la tua positività, ogni difficoltà ha trovato soluzione. Questo traguardo è anche merito tuo.

Grazie a tutte le altre colleghe universitarie tra cui Floarea, Silvia, Sara ed Elettra per la vostra gentilezza e solarità.

Infine, un segno di riconoscenza va anche al *team* della palestra *Ohana*, ormai la mia seconda casa, e in particolar modo al mio *coach* Leonardo che con estrema attenzione e precisione continua a trasmettermi questa grande passione.