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

Scienze Statistiche

Studio dell'evoluzione di una rete sociale: un'analisi tramite modelli bayesiani per reti

Relatore: Prof. Emanuele Aliverti
Dipartimento di Scienze Statistiche

Laureando: Virginia Murru
Matricola N. 2055493

Anno Accademico 2022/2023

Indice

Introduzione

Le reti sociali sono un tipo di dato complesso che permette di rappresentare in modo intuitivo ed efficace le relazioni tra entità. Questo tipo di struttura si focalizza sul legame presente tra gli individui e sui pattern e implicazioni di queste relazioni. Infatti, nelle reti sociali l'unità in analisi non è più l'individuo, ma le relazioni presenti tra gli individui. In generale, vengono chiamati *dati di rete* quei dati che riguardano entità interconnesse tra di loro e che misurano qualche tipo di relazione tra di esse. I dati di rete sono in grado di rappresentare in modo naturale sistemi complessi come ambienti sociali, politici ed economici che presentano una forte struttura di interdipendenza.

L'analisi delle reti sociali si pone l'obiettivo di modellare queste connessioni, nel tentativo di spiegare il meccanismo sottostante al comportamento degli individui, mentre le variabili esplicative proprie dei soggetti vengono messe in secondo piano. Infatti, l'obiettivo non è individuare il tipo di relazione presente tra variabile risposta e variabili esplicative, ma modellare le relazioni presenti tra gli individui.

L'analisi delle reti sociali può essere usata per rispondere a quesiti di tipo descrittivo e inferenziale. Le analisi di tipo descrittivo possono riguardare lo studio della struttura della rete o di alcune sue caratteristiche senza fare assunzioni distributive, mentre gli obiettivi inferenziali possono riguardare la bontà d'adattamento di un modello alla rete osservata, il confronto dell'andamento di più modelli e analisi di natura previsiva per gli archi della rete, o per un nuovo soggetto.

Questo tipo di struttura di dato ha anche il vantaggio di poter essere misurata nel tempo: è possibile monitorare la co-evoluzione delle connessioni tra gli stessi soggetti in più istanti temporali discreti. In questo caso si parla di reti sociali dinamiche, per le quali diventa d'interesse non solo analizzare i meccanismi che sottendono il comportamento degli individui, ma anche l'evoluzione temporale delle relazioni, ricercare entro le reti strutture di comunità e infine prevedere le connessioni presenti entro una rete per un istante di tempo futuro.

In questo elaborato di tesi verrà analizzata una rete sociale dinamica, che in seguito verrà chiamata *Social Evolution*, nella quale vengono misurate settimanalmente le connessioni tramite Bluetooth tra i telefoni dei partecipanti, che possono essere usate per misurare la vicinanza fisica dei soggetti. Tali

connessioni verranno modellate attraverso l'uso di variabili latenti tempo-dipendenti con l'ausilio dei modelli a spazi latenti e dei *latent position cluster models* con un numero di gruppi considerato fisso entro tutte le reti rilevate.

Nel primo capitolo verrà presentata una breve panoramica della genesi dei dati di rete, verrà poi definita in termini formali la struttura del dato di rete, la presentazione del dataset *Social Evolution* e le analisi preliminari effettuate, fino alla creazione del dataset nel suo formato finale. Infine, verrà presentata l'analisi esplorativa del dataset.

Nel secondo capitolo si offre una descrizione dei modelli utilizzati per l'analisi delle reti, formalizzandone la struttura, le assunzioni e le metodologie d'inferenza applicate.

Nel terzo capitolo si offre una prima applicazione dei modelli a dei dati sintetici simulati *ad hoc*, in modo da confrontare le loro prestazioni conoscendo il vero processo generatore dei dati.

Nel quarto capitolo si affronterà l'applicazione dei modelli proposti ai dati descritti in precedenza, verificando se i modelli proposti sono in grado di cogliere delle regolarità entro i dati considerati, si valuteranno quindi le prestazioni dei modelli sia in termini di replicabilità delle connessioni presenti entro la rete, sia nella previsione dei legami entro una rete ad un tempo futuro.

Capitolo 1

Definizione di dato di rete e problema in analisi

1.1 Genesi della ricerca sui dati di rete

L'interesse verso la rappresentazione di un sistema complesso attraverso la struttura dei dati di rete vede le sue origini nel decennio del 1930, con gli studi di Moreno (1934), Lewin (1936), Heider (1946). É stata poi dedicata un'importanza sempre maggiore, fino ad un aumento significativo negli anni '90 alla ricerca nell'ambito dell'analisi dei dati di rete, proseguito da Wasserman & Faust (1994), Freeman (2004). Questo interesse crescente origina dal fatto che molti fenomeni possono essere rappresentati attraverso un grafo, come interazioni tra elementi biologici (rapporti tra proteine), incidenze epidemiologiche riguardanti il contagio di malattie, interazioni tra strutture tecnologiche come collegamenti tra aeroporti o stazioni ferroviarie. Inoltre, con l'avvento dei *social network* un elevato numero di reti sociali, ovvero reti che riguardano persone, è diventato disponibile. Di conseguenza, è stato possibile studiare con un approccio scientifico questi dati, che forniscono informazioni sul modo in cui le persone si relazionano e sull'evoluzione dei loro legami. Oltre all'interesse in ambito sociologico di queste indagini, anche ai fini di marketing può essere interessante comprendere meccanismi di decisione umani.

Nel contesto delle reti sociali una delle prime proprietà a cui è stata dedicata particolare attenzione è quella dello *small world* da parte del sociologo Milgram (*The small world problem*, Milgram (1967)), il quale afferma che sono necessarie al massimo sei connessioni per raggiungere qualsiasi altro nodo presente nella rete (*Six degrees of separation*, originariamente teorizzato da Karinthy (1929) come 'gioco'). Come conseguenza, sin dagli albori dell'analisi dei dati di rete è stato d'interesse l'individuazione di nodi più connessi di altri, che possono assumere ruoli con significati diversi a seconda del contesto. Il concetto di 'ruolo' inteso come 'posizione sociale' che si riflette nella posizione spaziale entro la rete è stato introdotto da Lorrain & White (1971), la cui teoria si basa sulla definizione di *equivalenza*

strutturale per i nodi, secondo cui i nodi strutturalmente equivalenti assumono posizioni spaziali vicine nella rete; questo studio è stato poi esteso proponendo definizioni diverse di *equivalenza strutturale* di nodi, che hanno portato a soluzioni diverse, ma sempre basandosi sulla definizione di posizione sociale come collezione di attori coinvolti in modo simile in connessioni con altri attori. Il concetto di posizione è quindi distinto da quello di centralità del nodo, che è invece legato alla velocità di interazione con gli altri nodi (Bavelas, 1948); i nodi centrali sono quindi quelli che ricevono per primi un'informazione, per poi comunicarla agli altri nodi con cui sono collegati. Successivamente, quando è stato sviluppato il concetto di distanza geodesica, il concetto di centralità è stato associato a quello di 'distanza minima'. Si consideri per esempio le reti di criminalità organizzata o di qualche movimento sociale o politico d'interesse, l'individuazione dei soggetti più influenti consente di individuare i principali esponenti del movimento e le persone più vicine a loro. Nelle reti tecnologiche per esempio può essere d'interesse capire quali sono i nodi di spostamento più importanti e quali sono più marginali.

Lo studio di questi concetti ed in generale la ricerca relativa ai dati di rete è un campo in costante evoluzione ancora oggi, in cerca di approcci innovativi per comprendere e spiegare le strutture sottostanti ai sistemi complessi. Per ulteriori approfondimenti rispetto all'evoluzione di questo campo scientifico si può fare riferimento a Kilduff et al. (2023) ed a Kolaczyk & Csárdi (2014), cap. 1.

1.2 Definizione di rete dinamica

Una prima definizione di dato di rete è 'una collezione di nodi interconnessi da archi'. Le relazioni presenti tra i nodi possono essere di tipo diretto o indiretto. Il tipo di relazione può dipendere dalla natura dei dati, oppure può derivare da delle assunzioni fatte in fase di raccolta dei dati.

Il fenomeno che verrà analizzato in questo elaborato considera interazioni di tipo indiretto, quindi ogni arco può considerarsi reciproco. Per formalizzare il problema ed i dati considerati si può fare riferimento alla teoria dei grafi (Kolaczyk & Csárdi, 2014), che fornisce una definizione univoca della struttura del dato di rete. In forma generale, i termini 'rete' e 'grafo' possono essere considerati intercambiabili. Tuttavia, in modo più preciso, il termine 'rete' si riferisce al concetto generale, mentre il termine 'grafo' indica la struttura matematica sottostante alla rete.

Un grafo $G = (V, E)$ è una struttura matematica costituita da un insieme V di vertici, detti anche nodi, e da un insieme di archi E, definiti come una coppia $\{u, v\}$ di vertici distinti, $u, v \in V$. Con N_v si indica il numero di nodi presenti entro il grafo $N_v = |V|$. Secondo questa definizione, un grafo non contiene archi che collegano il nodo con se stesso, detti anche *loops*, ed un solo arco può interconnettere due nodi. Se un grafo possiede una di queste due proprietà, questo viene chiamato *multi-grafo*, in caso

contrario è denominato *grafo semplice*. Se gli archi entro il grafo presentano un ordinamento, il grafo viene chiamato *grafo diretto*, altrimenti viene chiamato *grafo indiretto*.

Le connessioni presenti entro un grafo indiretto possono essere rappresentate da una matrice $N_v \times N_v$ binaria simmetrica $\mathbf{A}$, detta matrice di adiacenza, o nel contesto delle reti sociali questa è anche detta 'socio-matrice', le cui componenti sono definite come:

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

Gli elementi diagonali A_{ii} della socio-matrice possono essere arbitrariamente posti come valori mancanti o possono essere fissati a zero.

Una possibile estensione della rete statica è costituita dalla rete dinamica, per cui uno stesso fenomeno relazionale può essere osservato in più istanti temporali. É possibile disporre di T reti che rilevano l'esistenza dello stesso tipo di relazione in T tempi, ponendosi sotto la restrizione di considerare sempre gli stessi vertici per tutti gli istanti temporali considerati. Si avranno a disposizione quindi T grafi e di T matrici di adiacenza, che possono essere rappresentate in modo compatto attraverso lo strumento matematico del tensore (*array*) a tre dimensioni $N_v \times N_v \times T$.

1.3 Statistiche descrittive di rete e proprietà

Per descrivere le caratteristiche predominanti a livello locale dei vertici della rete è necessario innanzitutto definire una misura di distanza tra i vertici di un grafo. Un modo per misurare la distanza tra due nodi è dato dalla lunghezza del percorso minimo. Tale percorso minimo, o *shortest path*, è definito come il numero minimo di archi necessari per interconnettere due nodi.
Per evidenziare le principali proprietà dell'intera rete si può usufruire di alcune statistiche descrittive, che possono essere suddivise in due categorie: statistiche *di nodo* e statistiche *di rete*. A livello locale si fa principalmente riferimento al grado e alla *betweenness* del nodo.

- il *grado* di un nodo i è definito come il numero di nodi con cui è connesso $d_i = \sum_{j=1}^{N_v} A_{ij}$;
- il livello di *betweenness* del nodo i equivale alla somma di tutti i nodi u e v in V diversi da i del rapporto tra il numero degli *shortest path* da u e v che passano per il nodo i $n_{uv}(i)$ e il numero totale di *shorterst path* tra u e v n_{uv}: $f_i = \sum_{u \neq i \neq v} \frac{n_{uv}(i)}{v_{uv}}$.

I tratti distintivi di una rete nel suo complesso si possono evincere considerando le statistiche della densità della rete, il numero di triangoli (o in generale di cicli) presenti e la transitività della rete.

- la densità del grafo è pari al rapporto tra il numero di archi osservati rispetto al numero di tutti gli archi possibili che si possono creare. Se la sociomatrice A è piena si ha $D = \frac{|E|}{|V|(|V|-1)}$, se invece si ha una rete indiretta nel quale solo la matrice triangolare superiore o inferiore assume valori allora $D = \frac{|E|}{\frac{|V|(|V|-1)}{2}}$;
- il numero di triangoli è calcolato come $T = \frac{tr(A^3)}{6}$;
- il livello di transitività entro una rete può essere calcolato come il rapporto tra il numero di triangoli presenti e il numero di *triplette*, che rappresenta il fenomeno di tre nodi collegati da due archi, senza considerare in modo stringente solo il caso di tre nodi tali per cui $i-j, j-k, k-i$. Una tripletta può invece essere formata da $i-j, j-k$. La transitività può anche essere calcolata attraverso il *global clustering coefficient* $C = \frac{6T}{\sum_{u \in V} d_u(d_u-1)}$, dove T rappresenta il numero di triangoli contenuti nel grafo;

Per un'estesa spiegazione di tali concetti si può fare riferimento a Wasserman & Faust (1994).
Avendo a disposizione una rete dinamica misurata in diversi istanti temporali, queste statistiche possono essere calcolate per ogni istante temporale, e da queste si possono trarre informazioni sia rispetto alla struttura della singola rete statica misurata in un tempo t, ma si può anche evincere l'evoluzione della struttura della rete dinamica osservando congiuntamente queste statistiche.
Come detto prima, le statistiche descrittive possono riassumere ed evidenziare alcuni aspetti presenti nella rete. Queste caratteristiche si traducono in proprietà possedute dalla rete, che sono comunemente oggetto di indagine e studio. Alcune di queste proprietà sono:

- *Small world*, come spiegato anche nella sezione 1.1, rappresenta il fenomeno per cui sono sufficienti pochi passaggi di nodi per raggiungere un qualsiasi nodo entro la rete;
- Transitività: propensione dei nodi a formare triangoli entro la rete. Considerando due nodi interconnessi con uno stesso nodo, è verosimile che anche questi due nodi siano interconnessi tra di loro;
- Omofilia per attributi: proprietà per cui i nodi che presentano alcune caratteristiche in comune hanno maggiore probabilità di essere interconnessi;
- Strutture di comunità: propensione dei nodi a formare gruppi con molte connessioni al loro interno e meno con gli altri nodi. Questa propensione è difficilmente catturata dalle variabili esplicative che possono essere raccolte;

- *Hubs*: nodi che presentano un elevato numero di connessioni e che fungono da tramite tra diversi vertici;

Il contesto dei dati di rete si distingue dall'approccio classico all'inferenza poiché la proprietà di omofilia per attributi, che coinvolge delle variabili esplicative a livello di nodo o di rete, non è più l'unico aspetto su cui ci si concentra, ma si colloca tra altre proprietà delle reti. Infatti, le variabili esplicative considerate spesso sono in grado di spiegare solo in piccola parte il meccanismo che porta i vertici a interconnettersi. La propensione alla creazione di relazioni tra i vertici è governata da diversi fattori, che comportano la presenza di forte dipendenza entro i dati.
Le proprietà appena presentate quali omofilia, transitività, *hubs* evidenziano una potenziale complessa gerarchia presente tra i nodi della stessa rete. Inoltre, questa forte gerarchia presente tra i nodi, che è dovuta a numerosi aspetti e che difficilmente è riconducibile ad un preciso schema, può suggerire quanto la struttura del dato di rete rappresenti un fenomeno complesso. Riuscire a spiegare le relazioni tra gli individui può essere più o meno complesso a seconda del fenomeno che descrive la rete.

1.4 Dataset di riferimento

I dati analizzati in questo elaborato sono stati raccolti da Madan et al. (2011). L'esperimento svolto dagli autori aveva lo scopo di studiare la co-evoluzione delle relazioni presenti entro una grande comunità, nel caso considerato alcuni studenti del MIT residenti nello stesso dormitorio. Il periodo di osservazione dei soggetti è intercorso tra Ottobre 2008 e Giugno 2009. Entro questo intervallo temporale i telefoni dei soggetti sono stati monitorati, e sono state raccolte informazioni rispetto alle chiamate da loro effettuate ed ai messaggi mandati; è stata anche monitorata la vicinanza dei telefoni dei soggetti attraverso le connessioni Bluetooth, che in seguito verrà chiamata 'prossimità'. Tale variabile rappresenta l'avvenuta ricezione (entro un raggio di 10 metri) di un segnale Bluetooth riscontrato tra due cellulari. Tale interazione viene registrata ad un preciso istante temporale. Oltre a queste variabili, sono anche disponibili informazioni rispetto al piano nel quale ogni studente vive e quale anno accademico frequenta.

L'intera analisi si incentra sulla modellazione della vicinanza (prossimità) tra i soggetti; le variabili riguardanti il numero di chiamate e messaggi effettuati verranno usate come variabili esplicative tempo-dipendenti, mentre le informazioni proprie di ogni nodo (piano di residenza, anno frequentato) verranno considerate entro la modellazione come variabili esplicative statiche.

1.4.1 Analisi preliminari

A partire dal 1 Ottobre 2008 sono state registrate stabilmente, ad intervalli di 6 minuti, le prossimità rilevate da un telefono ad un altro, con annesso *user ID* e *user ID receiver* in modo da poter collegare correttamente i soggetti. Le rilevazioni disponibili terminano il 26 Giugno 2009. Al fine di ottenere delle reti non eccessivamente sparse da analizzare, queste connessioni rilevate nel tempo sono state raggruppate in intervalli settimanali, tuttavia ogni rete è stata comunque trattata come binaria e non come rete pesata. A seguito di tale scelta, è stata monitorata nel tempo la presenza di almeno un contatto tra i soggetti partecipanti allo studio (inteso come interconnessione di un telefono con un altro via Bluetooth) entro un arco di tempo settimanale. Per non incorrere in distorsioni legate alla possibile diversa attività dei soggetti tra giorni feriali e festivi, l'analisi svolta parte dal 6 Ottobre 2008 fino al 26 Giugno 2009, in modo da considerare per ogni tempo una settimana completa dal lunedì alla domenica, avendo quindi intervalli equiripartiti. I nodi inclusi nella rete dinamica sono solo quelli risultati attivi in almeno un'istante temporale dei 37 considerati, che ammontano a 63 dagli 84 inclusi inizialmente nello studio. In conclusione, sono state ricavate 37 reti binarie con archi indiretti tra 63 nodi. Il numero di soggetti è stato mantenuto costante per tutta la durata dell'analisi temporale. Gli archi sono stati considerati indiretti seppur il tipo di informazione originariamente fornita sia di natura diretta, ma ai fini dell'analisi è ragionevole considerare la metrica 'vicinanza dei due telefoni'. Le variabili esplicative tempo-dipendenti che riguardano l'interazione tra i soggetti costituiscono a loro volta delle reti, che però risultano essere più sparse rispetto alla rete d'interesse, per cui è stato considerato un arco temporale maggiore, pari a due settimane, per non dover analizzare reti troppo sparse. Questo implica che per un istante temporale ogni due queste reti verranno mantenute costanti. Infine, sono state considerate anche le informazioni relative al piano nel quale gli studenti allogiano e l'anno accademico frequentato, non solo come variabili esplicative del singolo soggetto, ma anche al fine di considerare possibili effetti di convivenza nello stesso piano tra due soggetti e frequentazione del medesimo anno accademico.

1.5 Analisi esplorativa

Ai fini dell'analisi, è di primaria importanza verificare se ci sia variabilità tra le reti osservate ad ogni istante di tempo, tale da giustificare una modellazione di natura temporale. Vengono quindi mostrate in Figura 1.1, a scopo illustrativo, le reti al primo, al decimo, al ventesimo e all'ultimo tempo rilevato, le quali corrispondono alle prossimità degli studenti in quattro periodi temporali: a inizio anno scola-

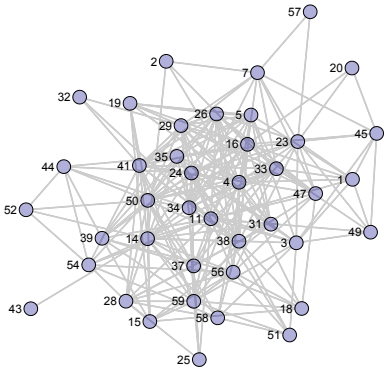
stico, a inizio Dicembre, a metà Febbraio, e infine a metà Giugno.

Nel contesto dei dati di rete la posizione spaziale dei nodi assume un ruolo molto importante, poiché fornisce informazioni rispetto alla popolarità del nodo, alla sua socialità, in relazione con quanto detto nelle sezioni precedenti. Infatti, esistono molteplici algoritmi che ottimizzano la posizione spaziale dei nodi secondo qualche criterio. Per le reti mostrate in Figura 1.1 è stato considerato un algoritmo di *Force Directed Placement*. I principali algoritmi di *Force Directed Placement* sono quelli di Kamada et al. (1989), e una sua versione più avanzata implementata da (Fruchterman & Reingold, 1991).

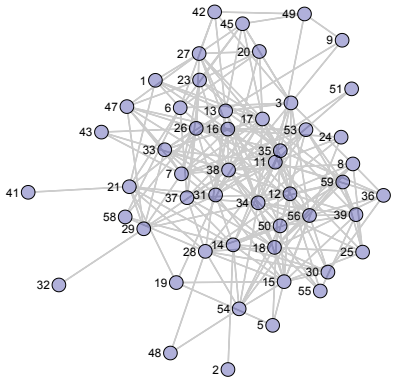
Osservando la Figura 1.1 si può vedere come il numero di nodi mostrati nelle quattro figure non sia sempre lo stesso, poiché questa rete dinamica è caratterizzata, per ogni tempo, da dei nodi che non presentano connessioni, che non sono sempre gli stessi, ma variano da tempo a tempo. Inoltre, si può vedere come il numero di archi e conformazione della rete evolva nel tempo: nel primo istante temporale rilevato sembrano esserci molti nodi aventi una posizione marginale nella rete e che presentano poche relazioni; nella rete al tempo 10 il numero di nodi aventi posizione marginale è inferiore, tuttavia si evidenziano vertici con una posizione ancora più estrema che nella rete precedente. Al tempo 20 la rete presenta una densità molto maggiore rispetto a tutti gli altri casi, quasi tutti i nodi sono abbastanza vicini alla rete. Nell'ultimo tempo considerato, il numero di archi è strettamente ridotto, sono distinguibili due comunità distinte entro la rete.

Per approfondire le differenze in termini descrittivi tra le reti osservate è possibile usare le metriche di riferimento descritte nella sezione precedente. Queste statistiche possono mettere in evidenza aspetti diversi delle reti a disposizione rispetto alla loro struttura: i gradi del nodo evidenziano in termini assoluti quante connessioni ogni nodo presenta, per cui si può vedere in modo generale se la rete ha un numero di nodi simile per la maggior parte dei nodi o se ha una struttura più disomogenea; la *betweenness* rappresenta il livello di centralità del nodo, per cui verosimilmente pochi nodi avranno valori alti, la maggior parte dei nodi avrà valori medi inferiori. I nodi che presentano valori elevati, in questo contesto magari per più istanti temporali, assumono il ruolo di *hubs* della rete, e, come anticipato prima, possono rappresentare il punto di tramite tra più gruppi. Il diametro della rete può dare un'indicazione rispetto alla sua forma, se più circolare (molti contatti tra tutti i nodi, eventualmente struttura disassortativa) o più allungata (forte attitudine al raggruppamento, struttura assortativa con comunità interconnesse da *hubs*); la transitività della rete e il numero di triangoli entro la rete forniscono informazioni rispetto alla presenza e alla 'forza' di strutture di comunità contenute al suo interno; la densità della rete riporta informazioni rispetto alla sparsità della stessa, ha dunque una più spiccata capacità riassuntiva.

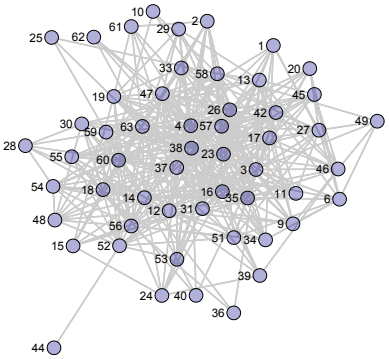
In Figura 1.2 sono mostrati i boxplot dei gradi e delle *betweenness* assunti da tutti i nodi ad



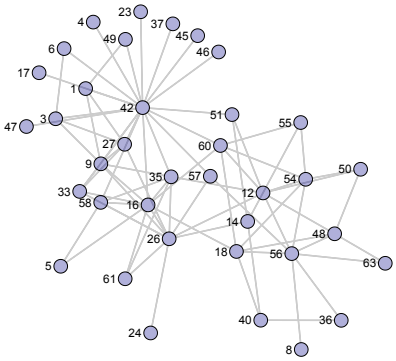
(a) Rete osservata al tempo 1.



(b) Rete osservata al tempo 10.



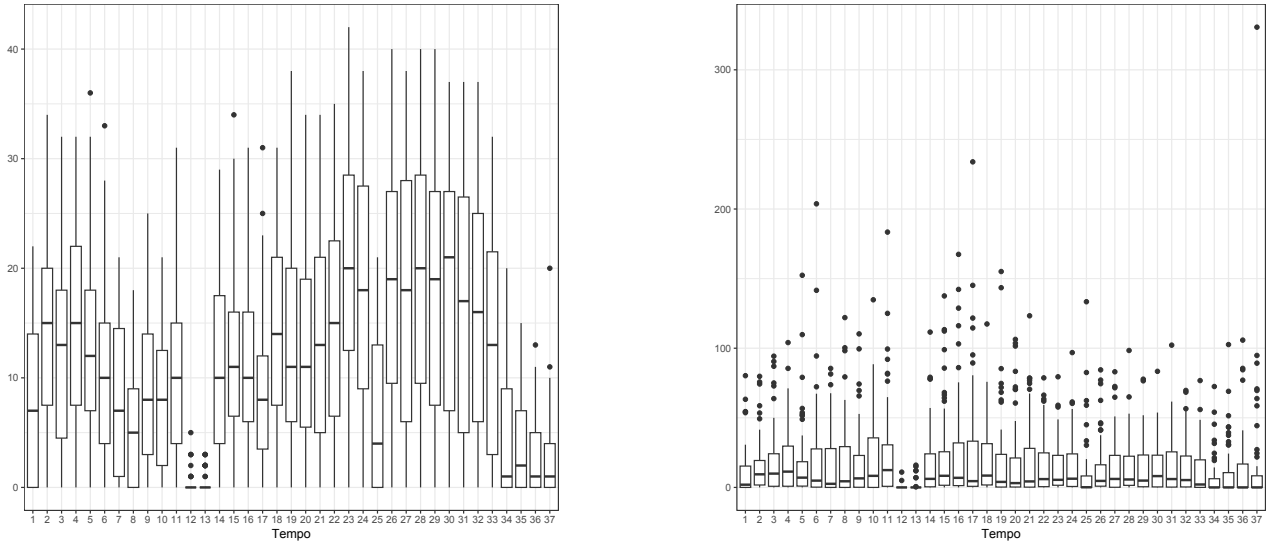
(c) Rete osservata al tempo 20.



(d) Rete osservata al tempo 37.

Figura 1.1: Rappresentazione grafica di quattro reti incluse nella rete dinamica corrispondenti a diversi istanti temporali.

ogni istante temporale. In questo modo è possibile evidenziare l'andamento globale della rete statica osservata ad ogni tempo, evidenziando eventuali differenze tra tempi. In Figura 1.2a vengono mostrati gli andamenti del grado dei nodi della rete, da questi ci si aspetta di ottenere distribuzioni con elevata variabilità, in relazione al fatto che la maggior parte dei nodi avranno un andamento simile, mentre una proporzione minore assumerà valori molto elevati e un'altra piccola parte valori molto ridotti. Tale comportamento è infatti rispettato, poiché il 75% dei vertici si aggira sempre su valori modesti,



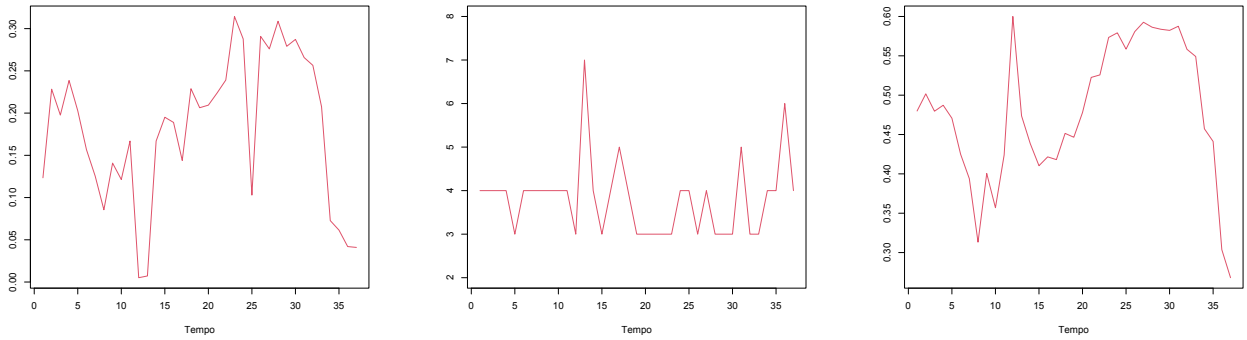
(a) Distribuzione dei gradi dei nodi.

(b) Distribuzione delle *betweenness* dei nodi.

Figura 1.2: Boxplot dei gradi e delle *betweenness* di tutti i nodi inclusi nello studio, in cui ogni boxplot si riferisce ad un tempo diverso.

mentre la restante parte è più densamente collegata, si evidenzia una pesante coda a destra per tutti i tempi.

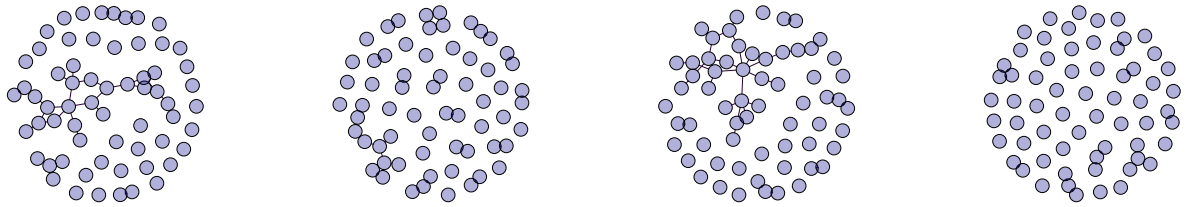
Osservando l'immagine mostrata in Figura 1.2a si possono notare alcune distribuzioni particolarmente schiacciate verso lo zero: innanzitutto si manifesta un calo del numero di relazioni tra i soggetti al tempo 12 e 13, ovvero dal 15 dicembre al 28 Dicembre, che corrisponde quindi al periodo di Natale, un altro al tempo 23, che corrisponde alla settimana tra il 9 Marzo 2009 e il 15 Marzo 2009, e infine si registra un andamento decrescente nelle ultime settimane del periodo di osservazione, compatibili con la fine dell'anno accademico. Questo fenomeno si riflette anche sulle altre metriche considerate. Si nota infine un aumento medio del numero di archi a partire dai tempi 18 fino al tempo 32. Osservando invece i boxplot della *betweenness* in Figura 1.2b si possono notare delle differenze di andamento, che però non sono così evidenti come quelle in Figura 1.2a, fatta eccezione per il periodo di fine Dicembre. L'andamento delle distribuzioni è comunque coerente rispetto alla teoria delle reti, per cui una piccola parte dei nodi assume dei valori particolarmente elevati, questi saranno i nodi più centrali della rete, mentre la maggior parte di essi assume valori nella media. Il numero di valori molto elevati sembra tendere ad aumentare in corrispondenza delle prime settimane di osservazione. Questo fenomeno potrebbe essere spiegato dal fatto che gli studenti non presentano tante connessioni tra di loro, per cui i nodi 'popolari' sono gli unici punti di tramite per alcuni percorsi tra nodi, mentre nella seconda metà i nodi presentano globalmente un numero di connessioni maggiore, e quindi i nodi centrali non hanno quel ruolo di centralità che avevano prima.



(a) Densità della rete dinamica. (b) Diametro della rete dinamica. (c) Transitività della rete dinamica.

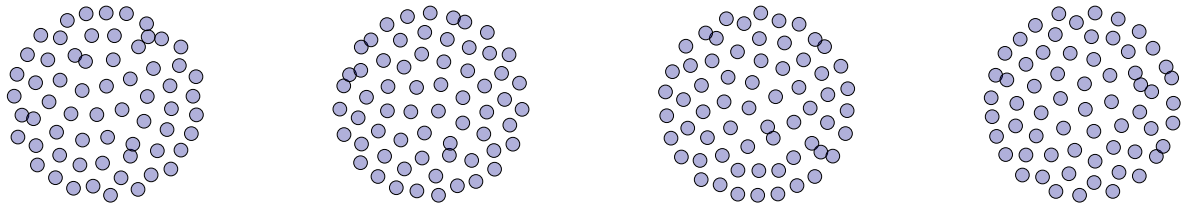
Figura 1.3: Statistiche descrittive a livello di rete. Poiché il tempo è discreto, vengono rappresentati dei valori puntuali per ogni tempo collegati da rette.

In Figura 1.3 viene mostrato l'andamento complessivo delle reti ad ogni tempo. In figura 1.3a viene mostrata la densità delle reti statiche ad ogni tempo, che assume dei valori modesti nella prima parte, aggirandosi sul 15% di connessioni presenti, rispetto a tutte quelle possibili, fino al periodo prima di Dicembre, ed aumenta in modo importante nella seconda parte del semestre, superando il 30% per alcuni tempi. La variabilità presente tra questi valori conferma e giustifica l'analisi della rete dinamica nel suo insieme, e inoltre evidenzia in modo più netto quanto osservato nella Figura 1.2a. La figura 1.3b mostra il diametro della rete rilevato nei diversi istanti di tempo, dal quale si può evincere che i nodi sono ben collegati tra di loro: i valori riportati si interpretano come il numero di passaggi di archi necessari per arrivare da un nodo ad un qualsiasi altro della rete (per i nodi che presentano connessioni), riportando quasi sempre un valore inferiore a 5. La regola dei 6 gradi di libertà di Milgram (1967) è quindi quasi sempre rispettata. In media sono necessari circa 4 passaggi per arrivare in qualsiasi punto della rete fino al tempo 17, e dal tempo 18 in poi questo valore decresce verso 3, in accordo con un numero di connessioni tra nodi maggiore. In alcuni momenti temporali questo valore aumenta, compatibilmente con gli istanti temporali in cui la rete è meno densa; questo fenomeno trova riscontro anche nel grafico della densità della rete mostrato in Figura 1.3a, che in quegli istanti temporali è particolarmente sparsa. In Figura 1.3c nella seconda metà dei tempi disponibili gli studenti sembrano essere più propensi a formare comunità, poiché si registra un incremento del livello di transitività, calcolato attraverso il *clustering coefficient* (sezione 1.3). Come è lecito aspettarsi, gli andamenti delle tre figure sono in accordo tra di loro, anche se riescono ad evidenziare aspetti leggermente diversi del fenomeno. Emerge quindi che a partire dal tempo 14 le interazioni tra i nodi aumentino: i soggetti presentano in generale un numero di connessioni maggiore, e nel contempo vengono formate più strutture a triangolo, fenomeno che suggerisce la creazione di comunità.



(a) Tempo 1. (b) Tempo 10. (c) Tempo 20. (d) Tempo 37.

Figura 1.4: Rete che collega i soggetti che, nell'arco di due settimane, si sono telefonati almeno una volta.



(a) Tempo 1. (b) Tempo 10. (c) Tempo 20. (d) Tempo 37.

Figura 1.5: Rete che collega i soggetti che, nell'arco di due settimane, si sono scambiati almeno un messaggio.

Si offre ora una breve analisi esplorativa delle variabili esplicative disponibili per i soggetti dello studio. Le reti relative alle chiamate, (Figura 1.4), e ai messaggi, (Figura 1.5), scambiati tra gli studenti sono molto sparse, tuttavia essendo comunque una fonte d'informazione disponibile sembra ragionevole includerla entro le analisi, ed eventualmente eliminarla nel caso in cui non apportasse nessun contributo in fase di modellazione. Va inoltre sottolineato che è possibile che gli studenti si sentissero tramite altri canali di comunicazione, come Facebook o twitter, (nati rispettivamente nel 2004 e 2006, per cui in un periodo molto fiorente per queste applicazioni). Potrebbe quindi essere possibile che il numero di chiamate e messaggi non sia rappresentativo dell'avvenuta comunicazione tra gli studenti, tuttavia si potrebbe anche pensare che le conversazioni che avvengono tramite canali *social* indichino una relazione più superficiale, mentre quelle avvenute tramite chiamate e messaggi siano legate ad un rapporto più profondo tra i soggetti, quindi a maggior ragione può valere la pena considerare queste reti anche se sparse.

Si riportano ora le tabelle di frequenza dell'anno accademico frequentato dagli studenti (Tabella 1.1) e del settore di residenza all'interno del campus (Tabella 1.2). Per entrambe le variabili esplicative si hanno dei valori mancanti, che non sono stati imputati ma sono stati mantenuti come 'Sconosciuto'.

282.1	282.2	282.3	282.4	290.2	290.3	290.4	Sconosciuto
11	7	12	11	5	11	4	5

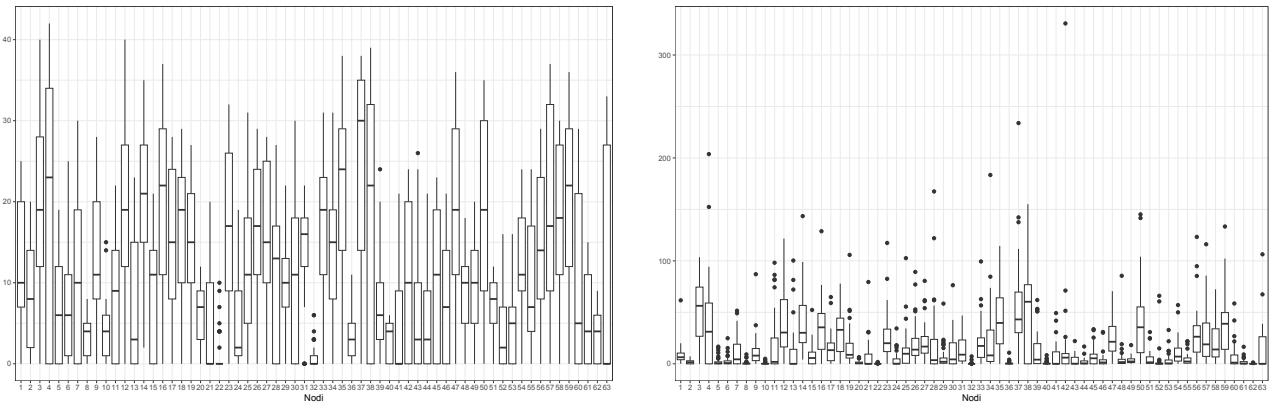
Tabella 1.1: Tabella di frequenza del piano di residenza degli studenti. Il primo numero corrisponde al numero identificato del palazzo, il secondo numero dopo il punto al piano di residenza.

Junior	Matricole	Secondo anno	Tutor Laureati	Senior	Sconosciuto
6	17	18	7	11	4

Tabella 1.2: Tabella di frequenza degli anni accademici frequentati dagli studenti.

La maggior parte dei soggetti entro lo studio è composta da studenti ai primi anni d'università, si registra infatti il 65% dei componenti dello studio tra Junior, ragazzi del primo e secondo anno. Un elevato numero di soggetti sembra provenire dal palazzo etichettato come 282, con una percentuale del 62%. Questo edificio sembra essere quello più frequentato dagli studenti ai primi anni, poiché il 73% degli studenti Junior, al primo e secondo ci risiede.

Per concludere l'analisi esplorativa viene mostrata la distribuzione dei gradi e della *betweenness* dei vertici non in funzione del tempo ma in funzione del vertice stesso, valutato su tutti gli istanti temporali disponibili (Figura 1.6). Questo consente di verificare visivamente quali siano i vertici più influenti o più attivi nei 37 istanti a disposizione. Si può ancora notare una forte disomogeneità tra nodi, che emerge sopratutto dal livello di *betweenness*, coerentemente con le proprietà delle reti. I nodi che presentano un maggiore livello di *betweenness* e grado medio sembrano essere il numero 3 e il 38. La variabilità presentata da queste distribuzioni non è però solo da attribuirsi ad una socialità dei



(a) Boxplot dei gradi dei nodi.

(b) Boxplot delle *betweenness* dei nodi.

Figura 1.6: Distribuzione dei gradi e delle *betweenness* dei nodi rispetto a tutti gli istanti temporali a disposizione, ogni boxplot si riferisce ad un nodo diverso. E' possibile vedere quali attori risultano essere più centrali (popolari) di altri.

vertici che può essere evoluta nel tempo, ma anche al fenomeno che si evince dalla Figura 1.1 per cui ad ogni istante temporale alcuni nodi non presentano connessioni, e questo fenomeno coinvolge nodi sempre diversi da tempo a tempo. Ci si aspetta allora che i nodi per cui tale fenomeno si presenta più spesso, cioè i nodi che presentano più 'assenze' dalla rete siano quelli le cui distribuzioni sono più variabili, che infatti presentano anche una coda sinistra molto accentuata.

Capitolo 2

Modelli utilizzati

In questo capitolo viene fornita una breve introduzione alle classi di modelli presenti in letteratura per dati di rete statici, verrà poi presentata una loro estensione per reti dinamiche. Nello specifico, verranno descritti i tre modelli applicati sia ad alcuni scenari di simulazione, sia al dataset *Social Evolution* precedentemente descritto. I primi due appartengono alla classe dei *Dynamic Latent Space Models*, il terzo è invece basato su una mistura di distribuzioni normali, che si colloca nel contesto dei *Latent Position Cluster Models*. In ogni caso, tutti i modelli considerati si rifanno all'uso di variabili latenti tempo-dipendenti che manterranno, per scelta di coerenza, la stessa interpretazione per tutto l'elaborato.

2.1 Contesto

I modelli presenti in letteratura per reti statiche possono essere divisi in due tipi di approcci: modelli marginali e modelli condizionali. I modelli marginali sono modelli rappresentati in termini di distribuzione congiunta di tutti gli archi e descrivono l'effetto dei predittori sui valori attesi marginali $\pi_{i,j}$; la classe di modelli più popolari che impone questa struttura è dato dai *Exponential Random Graph Models (ERGMs)*. I modelli condizionali assumono indipendenza tra coppie di nodi condizionatamente a delle variabili latenti specifiche per unità, che sono analoghe a degli effetti casuali. Queste variabili casuali vengono introdotte poiché si assume che i vertici possiedano delle caratteristiche che non sono state osservate o che non sono osservabili; per questi modelli si possono distinguere tre rami di sviluppo (Hunter et al., 2012): i modelli ad effetti casuali e ad effetti misti, i modelli a blocchi stocastici e i modelli a spazi latenti.

I modelli *ERGM* (Frank & Strauss, 1986) presentano una forte analogia con i modelli lineari generalizzati (*GLM*), legata al fatto che gli *ERGM* definiscono una distribuzione congiunta per tutti gli archi, e la distribuzione assunta per le probabilità di connessione presenta la forma di una famiglia esponenziale. Per questi modelli, Frank & Strauss (1986) hanno introdotto il concetto di *Dipendenza Markoviana*, secondo cui due archi sono dipendenti se presentano un nodo in comune, condizionatamente a tutti gli altri archi della rete. Questi modelli sono in grado di cogliere caratteristiche sia a livello di rete sia a livello di nodo, poiché è possibile incorporare entro il modello un elevato numero di predittori, quali predittori a livello di nodo, a livello di coppie di nodi e rispetto all'intera struttura della rete, non solo in termini di variabili esplicative ma di proprietà intrinseche dei nodi.

Il primo modello a variabili latenti introdotto è un modello ad effetti casuali sviluppato da van Duijn (1995). Esso può essere visto come un estensione del modello p_1 nel quale sono stati introdotti degli effetti casuali, e per questa ragione porta il nome di modello p_2. Il modello p_1 è uno dei primi modelli sviluppati per dati di rete, è stato introdotto da Holland & Leinhardt (1981) e rappresenta un caso specifico degli *ERGM*. van Duijn (1995) e van Duijn et al. (2004) hanno evidenziato il fatto che il modello p_1 possa essere visto come un modello lineare generalizzato, ed il modello p_2 come un modello lineare generalizzato ad effetti misti, che può essere stimato attraverso i minimi quadrati iterati generalizzati. Un metodo di inferenza bayesiana per il modello p_2 è stato sviluppato da Zijlstra et al. (2009).

I modelli volti alla ricerca di strutture di comunità entro la rete fanno riferimento ai *Stochastic Block Models*. In questa classe di modelli la variabile latente x_i rappresenta l'appartenenza del nodo i ad un gruppo latente, quindi $x_i = \{1, .., G\}, i = 1, \ldots, N_v$. Si assume che i vertici siano divisi in gruppi latenti e i membri dello stesso gruppo presentino caratteristiche simili. Più dettagliatamente, la probabilità di connessione tra due nodi è legata al gruppo d'appartenenza latente, si pone quindi l'attenzione sulle probabilità di connessione tra membri dello stesso gruppo e tra gruppi. Il primo modello che considera il numero di gruppi come ignoto è stato proposto da Nowicki & Snijders (2001).

I modelli a spazi latenti per reti sociali sono stati introdotti per la prima volta nel panorama letterario da Hoff et al. (2002), il quale tratta la modellazione di una rete statica sulla base di coordinate latenti in uno spazio a ridotta dimensionalità. In questo contesto, le variabili latenti x_i rappresentano una posizione distinta per ogni nodo $i = 1, \ldots, N_v$ in uno spazio latente, e assumono quindi un significato diverso rispetto a quelle introdotte negli *Stochastic Block-Models*. Lo spazio a ridotta dimensionalità nel quale si collocano le posizioni latenti è anche detto *social space*, e si riferisce ad uno spazio di caratteristiche latenti non osservate che rappresentano le tendenze di connessione potenziali entro la rete. Si assume che ogni nodo abbia una posizione ignota entro lo spazio sociale e si assume

inoltre che le probabilità di formazione degli archi tra i nodi, date le posizioni latenti associate, siano indipendenti. In particolare, vengono presentati due modelli che impongono una struttura diversa alle coordinate latenti: i modelli a distanze latenti (*latent distance models*) e i modelli di proiezione (*projection models*). La classe dei modelli a spazi latenti verrà utilizzata in fase di analisi, per cui queste due varianti verranno spiegate in modo dettagliato. Infatti, per analizzare i modelli che verranno utilizzati nel caso dinamico, occorre approfondire prima i modelli a spazi latenti che coinvolgono reti statiche, di cui i modelli per reti dinamiche rappresentano un'estensione.
Considerando archi di tipo indiretto, nel primo modello citato la probabilità di connessione tra nodi dipende dalla distanza euclidea tra le posizioni latenti ed è espressa come:

$$\begin{aligned} Y_{i,j} \mid \pi_{i,j} &\sim Bern(\pi_{i,j}), \\ \log\left(\frac{\pi_{i,j}}{1-\pi_{i,j}}\right) &= \alpha + \beta^T z_{i,j} - \|x_i - x_j\|. \end{aligned} \tag{2.1}$$

Dove $Y_{i,j} \in \{0,1\}$ rappresenta rispettivamente la presenza e l'assenza di un arco che interconnette il nodo i con il nodo j, $\pi_{i,j}$ rappresenta la probabilità di connessione tra i nodi i, j, $z_{i,j}$ indica il valore di una variabile esplicativa generica, che può essere diadica o relativa al nodo $z_{i,j} = z_i$. Al posto della distanza euclidea si può utilizzare qualsiasi misura di distanza che rispetti la disuguaglianza triangolare. La formula 2.1 impone simmetria tra la probabilità che il nodo i sia collegato con il nodo j e viceversa. Inoltre tiene intrinsecamente conto delle proprietà di reciprocità e transitività, e in presenza di variabili esplicative anche dell'omofilia. Nel secondo invece le variabili latenti sono rappresentate come punti in una sfera d-dimensionale di raggio unitario.

$$\begin{aligned} Y_{i,j} \mid \pi_{i,j} &\sim Bern(\pi_{i,j}), \\ \log\left(\frac{\pi_{i,j}}{1-\pi_{i,j}}\right) &= \alpha + \beta^T z_{i,j} + \frac{x_i^T x_j}{\|x_j\|}. \end{aligned} \tag{2.2}$$

Con l'equazione 2.2 la proprietà della transitività è ancora modellata, ma si tiene anche in considerazione il livello specifico di attività del nodo.

Questi due approcci si differenziano per l'uso della distanza euclidea nel primo modello e per l'uso del prodotto scalare nel secondo, queste risultano essere entrambe metriche simmetriche, per cui tengono conto di proprietà come la transitività, e quando sono presenti delle variabili esplicative rispettano la proprietà dell'omofolia. Come conseguenza della diversa struttura, le coordinate latenti assumono interpretazioni differenti nei due casi considerati: nel modello a distanze latenti, due nodi che presentano coordinate latenti molto vicine a livello spaziale avranno più elevata probabilità di essere interconnessi, mentre nel modello di proiezione due nodi che presentano coordinate latenti aventi stes-

sa direzione avranno maggiore probabilità di connessione. Questo può essere spiegato intuitivamente, considerando per esempio due dimensioni latenti, come da punti che si trovano nello stesso quadrante. Il modello a distanze latenti proposto da Hoff et al. (2002) è stato esteso al contesto delle reti dinamiche da Sarkar & Moore (2005), che è stato a sua volta esteso da Sewell & Chen (2015), formalizzando però un approccio statistico per l'analisi della rete, mentre il metodo proposto da Sarkar & Moore (2005) utilizza algoritmi di ottimizzazione per la stima. Nei modelli per reti dinamiche si suppone anche che queste posizioni possano cambiare ad ogni istante temporale, ma si ritiene improbabile che cambino in modo sostanziale.

La classe dei modelli a blocchi stocastici non è stata affrontata in questo elaborato di tesi in quanto si è preferito mantenere la precedente rappresentazione e significato delle variabili latenti, come anticipato nell'Introduzione del presente capitolo. Per ricercare strutture di comunità entro la rete mantenendo la stessa interpretazione delle variabili latenti è stato utilizzato un modello a distanze latenti con mistura, sviluppato da Krivitsky & Handcock (2008). Considerando una rete formata da archi indiretti, il predittore lineare è lo stesso del modello a distanze latenti (Equazione 2.1), la grande differenza sta nell'assunzione della distribuzione a priori delle posizioni latenti. Mentre nei modelli a spazi latenti viene posta una distribuzione normale con una certa varianza per ogni componente delle posizioni latenti, in questo caso si assume

$$x_i \sim \sum_{g=1}^{G} \lambda_g N_d(\mu_g, \sigma_g^2 I). \tag{2.3}$$

Il modello mantiene dunque tutte le proprietà di quello a distanze latenti, ma si ha in aggiunta la possibilità di indagare sulle eventuali comunità presenti entro la rete.

2.2 Quantità di base

In tutti i contesti che verranno presentati di seguito, si lavorerà con una rete dinamica costituita da un *array* a tre dimensioni $y = (y_{i,j,t})_{i,j=1,\dots,N_v,t=1,\dots,T}$, $y_{i,j,t} \in \{0,1\}$, dove $y_{i,j,t}$ indica la presenza o assenza di un arco tra i vertici i e j al tempo t, con N_v si identifica il numero di nodi presenti nelle reti, e con T il numero di tempi discreti per cui si ha un'osservazione della rete. Occorre inoltre effettuare un'altra distinzione tra la variabile casuale $Y_{i,j,t}$ e la rete dinamica osservata $y_{i,j,t}$, intesa come realizzazione di $Y_{i,j,t}$. Si sottolinea che il numero di nodi presente nella rete dinamica è considerato fisso per tutti gli istanti temporali. Il tensore delle probabilità di connessione tra i vertici ad ogni istante temporale

$\pi = (\pi_1, \ldots, \pi_T), \pi_{i,j,t} \in (0,1), \dim(\pi) = N_v \times N_v \times T$, dove ogni elemento $\pi_{i,j,t}$ indica la probabilità di connessione tra il nodo i e il nodo j al tempo t, $i = 1, ..., N_v$ e $j = i+1, ..., N_v$. Sia Z l'array di dimensione $N_v \times p \times T$, contenente le variabili esplicative disponibili per ogni nodo. Queste possono essere tempo-dipendenti oppure invarianti temporalmente. Ad ogni modo, con questa scrittura Z è stato definito nel modo più generico possibile. Si considerano le coordinate latenti $\mathbf{x} \in \mathbb{R}^{N_v \times d \times T}$; nelle sezioni successive si farà riferimento a $\mathbf{x}_{i,t}$ inteso come vettore d-dimensionale delle posizioni latenti del nodo i al tempo t, la sua singola componente verrà indicata con $x_{i,h,t}$, mentre con $\mathbf{x}_t$ ci si riferirà alla matrice delle posizioni latenti di tutti i nodi corrispondenti ad un tempo t, di dimensioni $N_v \times d$. Si considera infine una variabile $K_i \in \{1, \ldots, G\}$, $i = 1, \ldots, N_v$ che rappresenta l'appartenenza del vertice i al gruppo $g \in \{1, \ldots, G\}$, dove G indica il numero di gruppi selezionati.

2.3 Modelli implementati

Come anticipato precedentemente, l'obiettivo è quello di spiegare tramite modellazione i meccanismi sottostanti alla creazione di legami tra vertici e i pattern di connessione, e spiegare anche i meccanismi che sottendono la loro evoluzione temporale.

Questo elaborato di tesi si concentra sull'implementazione di due modelli appartenenti alla classe dei modelli a spazi latenti dinamici, insieme ad una loro variante che contempla una distribuzione mistura. Questi approcci modellano le probabilità di connessione rispetto alla posizione degli attori nello spazio latente, che può variare nel tempo. Questo consente da una parte di indagare i cambiamenti strutturali che possono verificarsi nella rete nella sua evoluzione temporale, dall'altra riescono a cogliere correttamente il comportamento della rete e degli attori ad ogni istante temporale. Il primo modello che verrà trattato è stato sviluppato da Durante & Dunson (2014), nel quale ogni componente delle posizioni latenti si distribuisce secondo un processo gaussiano. Il secondo modello considerato è stato invece presentato da Sewell & Chen (2015), e assume una struttura autoregressiva del primo ordine per le posizioni latenti. Questi due algoritmi riescono a cogliere correttamente proprietà sia di nodo e sia delle reti e, con l'ausilio di variabili esplicative, riescono a catturare caratteristiche quali l'omofilia per attributi. L'ultimo modello che verrà trattato è un'estensione per reti dinamiche del *latent position cluster model* presentato da Handcock et al. (2007), il quale assume una distribuzione normale multivariata con mistura per le posizioni latenti e assume il numero di gruppi presenti entro la rete dinamica come ignoto. Si sottolinea che i tre modelli presentati includono le due formulazioni previste da Hoff et al. (2002) presentate in equazione 2.1 e 2.2. Per non complicare eccessivamente il modello, è stato posto il vincolo di appartenenza di ogni nodo ad un solo gruppo per tutti i tempi,

questa restrizione potrebbe risultare abbastanza forte a seconda della struttura della rete dinamica osservata. Inoltre, le posizioni latenti a priori presentano la stessa distribuzione per tutti i tempi, non viene quindi considerata una vera e propria evoluzione temporale come per i due modelli precedenti. Tuttavia le posizioni latenti possono comunque assumere un valore diverso per ogni tempo, mantenendo quindi una certa misura di flessibilità temporale. Il modello a distanze latenti con mistura presenta il vantaggio di riuscire a considerare, oltre alle proprietà appena citate, anche la naturale tendenza dei nodi a formare delle comunità, ma risulta avere dei vincoli un po' più forti rispetto agli altri due modelli in termini di evoluzione temporale.

2.4 Modello a proiezioni latenti dinamico

Sotto le condizioni esposte nella sezione 2.2, la distribuzione di probabilità delle componenti $Y_{i,j,t}$ della sociomatrice Y_t è pari a

$$Y_{i,j,t}|\pi_{i,j,t} \sim Ber(\pi_{i,j,t}), \tag{2.4}$$

indipendenti per ogni $i = 1, ..., N_v$ e $j = i + 1, ..., N_v$, dove $\pi_{i,j,t} = \pi_{j,i,t} = \mathbb{P}(Y_{i,j,t} = 1)$.
Il predittore lineare di $Y_{i,j,t}$ è definito come

$$\log\left(\frac{\pi_{i,j,t}}{1-\pi_{i,j,t}}\right) = \mu_t + \mathbf{x}_{i,t}^T\mathbf{x}_{j,t} + \beta^T Z_{i,j,t}. \tag{2.5}$$

Oltre alle posizioni latenti si considera un termine di intercetta μ_t anch'esso tempo-dipendente, ma unidimensionale, e un parametro β, che può essere definito come un vettore p-dimensionale, in forma del tutto generale, che non è però tempo-dipendente, ma bensì statico. La decomposizione $\mathbf{x}_t^T\mathbf{x}_t$ ha l'importante caratteristica di non essere unica, per cui le posizioni latenti non sono identificabili in modo univoco. L'assenza di identificabilità delle posizioni latenti deriva proprio dal fatto che le probabilità di connessione in equazione 2.5 dipendono dalle posizioni latenti solo attraverso la distanza tra coppie di posizioni individuali, per cui ogni rotazione applicata a $\mathbf{x}_t^T\mathbf{x}_t$ porta allo stesso valore della verosimiglianza. Poiché però in questo contesto si è interessati all'inferenza su π_t, non risulta necessario imporre dei vincoli di identificabilità su $\mathbf{x}_t$ poiché non precludono l'identificabilità di π_t, come riportato in Gosh & Dunson (2009). In ogni caso, è possibile utilizzare il *Procrustes Method* per avere una decomposizione univoca che minimizzi la distanza rispetto ad un set di coordinate predefinito, questo processo può comunque essere effettuato dopo aver simulato dalla distribuzione a posteriori tali coordinate, poiché le $\mathbf{x}^*$ ottenute con il *Procrustes Method* e $\mathbf{x}$ portano alle stesse conclusioni inferenziali.

Per imporre una distribuzione a priori sulla matrice delle probabilità di connessione π_t si impone una distribuzione su μ_t e sulla distribuzione delle coordinate latenti $\mathbf{x}_t$. Nello specifico, si considera un processo gaussiano per i fattori latenti tale che

$$\begin{aligned} x_{i,h,t} &\sim GP(0, \tau_h^2 c_X), \\ \{x_{i,h,1}, \dots, x_{i,h,T}\} &\sim N_T(0, \tau_h^2 \Sigma), \\ c_X(t_i, t_j) = \Sigma_{ij} &= \exp\{-\kappa_X (t_i - t_j)^2\}, \quad t_i, t_j \in \{1, \dots, T\}. \end{aligned}$$

I parametri τ_h^2 presentano distribuzione a priori pari a

$$\tau_h^2 \sim IGa(a, b), \;\; h = 1, \dots, d. \tag{2.6}$$

In questo modo si assume dipendenza temporale per le coordinate latenti relative alla dimensione latente h-esima, mentre si assume indipendenza tra coordinate latenti di nodi distinti e di dimensioni distinte.
In modo analogo si impone a priori un processo gaussiano per

$$\begin{aligned} \mu_t &\sim GP(0, c_\mu), \\ \{\mu_1, \dots, \mu_T\} &\sim N_T(0, \tilde{\Sigma}), \\ c_\mu(t_i, t_j) = \tilde{\Sigma}_{ij} &= exp\{-\kappa_\mu (t_i - t_j)^2\}, \quad t_i, t_j \in \{1, \dots, T\}. \end{aligned} \tag{2.7}$$

Le distribuzioni a priori 2.6 e 2.7 di tipo non parametrico forniscono molta flessibilità alle componenti. La dipendenza temporale viene imposta attraverso la matrice di varianza e covarianza Σ e $\tilde{\Sigma}$, ponendo quindi una correlazione diversa da zero per i tempi vicini rispetto a quello considerato, questa correlazione decresce all'aumentare della distanza temporale tra le variabili, tendendo a zero. Tali matrici sono molto sensibili rispetto al valore assunto dai parametri di scala κ_μ e κ_X, poiché regolano la 'forza' della dipendenza temporale imposta alle variabili. In questo modo le coordinate dei fattori latenti e la media globale hanno un'evoluzione in un tempo continuo.
Una distribuzione a priori per i parametri β può essere data dalla distribuzione normale multivariata. La previsione di una nuova rete al tempo $T + 1$, avendo a disposizione solo le reti per i tempi $\mathcal{T} = \{1, \dots, T\}$, si può effettuare tramite simulazione dalla distribuzione predittiva a posteriori. Se si è interessati alla rete $Y_{i,j,T+1}$ si possono simulare le posizioni latenti e i termini di intercetta per $T+1$ tempi, poiché la distribuzione imposta alle posizioni latenti è di tipo T-variato, per cui il termine di intercetta e le posizioni latenti per un tempo futuro vanno simulate insieme alle altre T. Dopo aver

tratto i campioni dalla distribuzione predittiva a posteriori, questi possono essere usati per calcolare in modo approssimato le proprietà della rete futura simulata.
Considerando il modello appena descritto senza variabili esplicative, è importante sottolineare l'interpretazione dei parametri inseriti entro al predittore lineare in equazione 2.5. Il coefficiente μ_t rappresenta la media globale del numero di connessioni presente entro la rete. Le coordinate latenti misurano la probabilità di connessione tra gli attori al netto del numero di archi presente entro la rete nell'istante temporale t-esimo. Questo tipo di struttura è in grado di catturare la dipendenza triadica entro i nodi della rete, strettamente legata al concetto di omofilia e di transitività. Come per il *projection model* di Hoff et al. (2002) se le coordinate latenti di due nodi hanno la stessa direzione, i due nodi avranno probabilità maggiore di essere interconnessi.

2.5 Modello a distanze latenti dinamico

Il modello nella sua formulazione originaria definita nell'articolo di Sewell & Chen (2015) considera una rete binaria con archi diretti, quindi verrà considerata una sua versione semplificata per conformarsi al problema in esame.
Ci si pone ancora sotto le condizioni esposte nella sezione 2.2. Le posizioni latenti si distribuiscono secondo un processo autoregressivo in cui, al primo istante temporale ha distribuzione iniziale

$$\mathbf{x}_{i,1} \mid \theta \sim N_d(0, \tau^2 I), \tag{2.8}$$

indipendente per ogni $i = 1, \ldots, N_v$. Per i restanti istanti temporali si ha l'equazione di transizione:

$$\mathbf{x}_{i,t} \mid \mathbf{x}_{i,t-1}, \theta \sim N_d(\mathbf{x}_{i,t-1}, \sigma^2 I), \qquad t = 2, \ldots, T.$$

Si suppone inoltre che le osservazioni $Y_{i,j,t}$, per uno stesso t fissato, siano indipendenti condizionatamente a $\mathbf{x}_t$ e al vettore dei parametri θ, che verrà definito dopo aver presentato tutte le quantità coinvolte. Tale forma di dipendenza si formalizza scrivendo:

$$Y_{i,j,t} \mid \mathbf{x}_{i,t}, \mathbf{x}_{j,t}, \theta \sim Ber(\pi_{i,j,t}),$$

In questo elaborato il modello proposto da Sewell & Chen (2015) è stato complicato introducendo un predittore lineare alternativo che considerasse anche un termine di intercetta variabile nel tempo e

presentasse l'inclusione di variabili esplicative. Si può allora riscrivere

$$\begin{aligned} &Y_{i,j,t} \mid \mathbf{x}_{i,t}, \mathbf{x}_{j,t}, \theta \sim Ber(\pi_{i,j,t}), \\ &\log\left(\frac{\pi_{i,j,t}}{1-\pi_{i,j,t}}\right) = \mu_t + \gamma(1 - \|\mathbf{x}_{i,t} - \mathbf{x}_{j,t}\|) + \beta^T Z_{i,j,t}. \end{aligned} \tag{2.9}$$

Il parametro multidimensionale θ è dato da $\theta = (\tau^2, \sigma^2, \beta, \gamma, \mu_1, \ldots, \mu_T)$, dove il coefficiente β è un vettore p-dimensionale, espresso in forma generale. La formula mantiene una scrittura comunque analoga rispetto a quella presente nel modello a distanze latenti, mostrato in equazione 2.1.
Si vuole quindi fare inferenza sulla distribuzione a posteriori $p(\mathbf{x}_1, \ldots, \mathbf{x}_t, \theta \mid Y_1, \ldots, Y_t)$. Come per il modello precedente, le posizioni latenti sono invarianti rispetto a rotazioni, portano ancora alla stessa verosimiglianza, perciò non sono identificabili. La distribuzione a posteriori sarà invariante rispetto a rotazioni, traslazioni e trasformazioni delle posizioni latenti. Si può anche in questo caso ricorrere al *Procrustes Method* per ricavare un set di posizioni latenti univoco, tale da minimizzare la distanza rispetto ad un set di coordinate scelto. Come per il modello precedente, quest'operazione può essere svolta dopo aver simulato dalla distribuzione a posteriori, poiché il predittore lineare rimane comunque invariato per qualsiasi rotazione delle posizioni latenti.
Vengono fissate delle distribuzioni a priori dei parametri θ pari a

$$\begin{aligned} \gamma &\sim N(\eta, \xi), \\ \sigma^2 &\sim IGa(a_\sigma, b_\sigma), \\ \tau^2 &\sim IGa(a_\tau, b_\tau), \\ (\mu_1 \ldots, \mu_T) &\sim N_T(0, \tilde{\Sigma}), \\ \beta &\sim N_p(0, \sigma^2_\beta I), \end{aligned}$$

supponendo ancora un processo gaussiano per la componente t-esima dell'intercetta, definito come in equazione 2.7.
Prevedere una nuova rete Y_{T+1} rappresenta nuovamente un interrogativo d'interesse, come anche prevedere le posizioni latenti $\mathbf{x}_{T+1}$. Per prevedere la futura evoluzione delle posizioni latenti si può simulare dalla distribuzione a posteriori delle posizioni latenti fino al tempo T, rispettando la struttura Markoviana imposta dal modello. Simulando $\widetilde{M}$ campioni dalla distribuzione a posteriori di θ, si può simulare dalla distribuzione predittiva a posteriori $\widetilde{Y}^{(m)}_{T+1} \sim p(Y_1, \ldots, Y_T \mid \theta^{(m)})$ per $m = 1, \ldots, \widetilde{M}$. Le proprietà empiriche della rete futura possono essere calcolate in modo approssimato sulla base dei campioni simulati. Questo approccio è analogo a quello descritto nella sezione precedente, differisce

solo per la modalità di simulazione di θ.

I coefficienti presenti nel predittore lineare mostrato in equazione 2.9 possono assumere la seguente interpretazione. Il parametro μ_t rappresenta il numero medio di connessioni entro la rete all'istante temporale t-esimo. Le probabilità di connessioni tra i nodi sono regolate dalle posizioni latenti, per cui punti vicini nello spazio corrisponderanno ad elevate probabilità. Il parametro γ può essere interpretato come un parametro di scala rispetto all'influenza delle posizioni latenti. In particolare, se questa distanza sarà inferiore a 1 si avrà un effetto positivo (ci si aspetta che γ sia positivo), all'aumentare della distanza si avrà una probabilità di connessione sempre minore. Data la presenza dell'intercetta, il parametro γ va interpretato come scostamento rispetto al numero di connessioni medio della rete al tempo t. Lo spazio latente può essere pensato come uno spazio delle caratteristiche nel quale le distanze tra gli attori riflettono quanto questi sono simili o la forza della loro relazione.

2.6 Modello a distanze latenti con mistura

La formulazione di questo modello, come anticipato nelle sezioni 2.1 e 2.3, è basata sul *latent position cluster model* sviluppato da Handcock et al. (2007), il quale tratta il caso di una rete statica e si concentra su una rete indiretta, per cui verrà proposto un ampliamento tale da consentire alle posizioni latenti di poter evolvere nel tempo. Inoltre, le distribuzioni a priori assunte per alcuni parametri si distanziano leggermente dalla formulazione originale. Nello specifico, si considerano quelle presentate nel pacchetto *latentnet* redatto da Krivitsky & Handcock (2008).
In questo modello la distribuzione delle posizioni latenti dipende dal gruppo di appartenenza del nodo, considerato fissato per tutto il periodo d'osservazione. Il numero di gruppi viene fissato a priori e costituisce quindi un parametro di regolazione.
Si considera ancora uno *social space* nel quale giaciono le variabili latenti, e inoltre si suppone ancora che gli archi $Y_{i,j,t}$ siano indipendenti condizionatamente alle posizioni latenti $\mathbf{x}_{i,t}$ e $\mathbf{x}_{j,t}$. Rispettando ancora la struttura esposta nella sezione 2.2, l'assunzione probabilistica può essere formalizzata come:

$$P(Y_t|Z_t, \mathbf{x}_t, \theta) = \prod_{i \neq j} P(Y_{i,j,t} \mid \mathbf{x}_{i,t}, \mathbf{x}_{i,t}, Z_{i,j,t}, \theta).$$

Il parametro θ contiene tutti i parametri del modello. Il predittore lineare è definito in modo analogo al *latent distance model* di Hoff et al. (2002).

$$\begin{aligned} Y_{i,j,t} \mid \mathbf{x}_t &\sim Ber(\pi_{i,j,t}), \\ \log\left(\frac{\pi_{i,j,t}}{1-\pi_{i,j,t}}\right) &= \mu_t + \gamma(1 - \|\mathbf{x}_{i,t} - \mathbf{x}_{j,t}\|) + \beta^T Z_{i,j,t}. \end{aligned} \tag{2.10}$$

La distribuzione delle posizioni latenti è quindi data da una mistura di distribuzioni normali multivariate del tipo

$$\mathbf{x}_{i,t} \sim \sum_{g=1}^{G} \lambda_g N_d(\upsilon_g, \sigma_g^2 I). \tag{2.11}$$

Il parametro λ_g corrisponde alla probabilità d'appartenenza al gruppo g. L'indipendenza tra le d componenti della distribuzione multivariata è giustificata dal fatto la verosimiglianza è invariante rispetto a rotazioni dello spazio sociale, per cui il modello viene definito in modo indipendente dal sistema di coordinate.

Il numero di dimensioni latenti d per la distribuzione delle posizioni latenti ed il numero di gruppi G sono i parametri di regolazione del modello, che andranno quindi scelti secondo qualche criterio. I metodi di selezione di questi parametri considerati in questo elaborato verranno presentati nella sezione 2.8.

Di seguito vengono presentate le distribuzioni a priori utilizzate.

$$\begin{aligned} \upsilon_g &\sim N_d(0, \omega^2 I_d), \quad g = 1, \ldots, G, \\ \sigma_g^2 &\sim \sigma_0^2 Inv\chi_\alpha^2, \quad g = 1, \ldots, G, \\ (\lambda_1, \ldots, \lambda_G) &\sim Dirichlet(\nu_1, \ldots, \nu_G). \end{aligned} \tag{2.12}$$

Per prevedere una rete futura, un metodo di semplice applicazione consiste nel simulare dalla densità predittiva a posteriori in modo analogo alla procedura spiegata per gli altri due metodi, nel quale l'unica differenza consiste nel simulare dalla distribuzione a posteriori le posizioni latenti, che però mantengono ancora l'assegnazione al gruppo g, per cui le posizioni latenti per un tempo $T+1$ possono essere simulate conoscendo da quale media e varianza simulare.

L'appartenenza al gruppo per ogni vertice K_i può essere calcolata come

$$\hat{K}_i = \underset{g=1,\ldots,G}{\operatorname{argmax}} \mathbb{P}(K_i = g) = \frac{\mathbb{P}(\mathbf{x}_{i,t} \mid \upsilon_g, \sigma_g^2)\, \mathbb{P}(K_i = g \mid \lambda)}{\sum_{g'=1}^{G} \mathbb{P}(\mathbf{x}_{i,t} \mid \upsilon_{g'}, \sigma_{g'}^2)\, \mathbb{P}(K_i = g' \mid \lambda)}. \tag{2.13}$$

Il gruppo d'appartenenza selezionato è quello per cui si ottiene una probabilità d'appartenenza massima. La formula in equazione 2.13 può essere calcolata per un qualsiasi tempo, poiché, come detto in precedenza, l'appartenenza al gruppo non varia tra i diversi istanti temporali.

L'interpretazione dei parametri del predittore lineare riportati in Equazione 2.10 è analoga a quella

del modello a distanze latenti dinamico (sezione 2.5), considerando in aggiunta il parametro λ_g, che corrisponde alla proporzione di nodi appartenenti al gruppo g-esimo.

2.7 Inferenza Variazionale

Il metodo cardine per fare inferenza su una distribuzione a posteriori è il metodo Monte Carlo, il quale prevede di simulare un elevato numero di valori dalla distribuzione a posteriori e stimare le quantità d'interesse tramite integrazione Monte Carlo. Tra questi metodi uno dei più utilizzati è il *Markov Chain Monte Carlo*, nel quale si usano le catene di Markov per simulare una serie di valori dipendenti provenienti dalla distribuzione a posteriori. Nonostante i valori simulati siano dipendenti, e abbiano quindi un contenuto informativo inferiore rispetto a delle simulazioni indipendenti, questi possono comunque essere usati per fare inferenza sulla distribuzione a posteriori. Il minor contenuto informativo legato alla dipendenza dei valori può essere facilmente ovviato generandone un numero maggiore. A seconda della complessità del modello, la catena può necessitare di un numero di iterazioni variabile prima di convergere alla distribuzione a posteriori. Il numero di iterazioni della catena può quindi essere elevato, e a seconda della complessità del problema può richiedere lunghi tempi d'esecuzione. Quindi, per ragioni legate alla scalabilità del problema in questione, è stato deciso di optare per un metodo alternativo al canonico MCMC, costituito dall'Inferenza Variazionale (Blei et al., 2017). La scelta di usare l'approccio dell'Inferenza Variazionale è legato alla elevata dimensione del dataset reale che si vuole analizzare, in particolare al fatto che il metodo MCMC ha un tempo di esecuzione fortemente legato al numero di istanti temporali della rete dinamica, mentre il numero di nodi non influenza il tempo di convergenza del metodo. Avendo quindi una rete dinamica costituita da 37 istanti temporali il metodo MCMC avrebbe comportato tempi di esecuzione molto lunghi. In termini quantitativi, il metodo MCMC ha richiesto circa 160 minuti per avere dei buoni risultati in termini di mixing, mentre il metodo VI utilizzato ha richiesto circa 180 secondi per ogni modello. I tempi richiesti dal primo metodo non sono quindi proibitivi ma risultano molto elevati. A fronte di un risultato approssimato, si è preferito utilizzare un metodo che presentasse un tempo d'esecuzione ridotto.

L'inferenza Variazionale, anziché usare l'idea del campionamento, propone una procedura di ottimizzazione. Il metodo VI si pone lo stesso obiettivo del metodo MCMC, ma usa un approccio diverso: il metodo MCMC campiona da una catena di Markov e approssima la distribuzione a posteriori con campioni tratti da essa, mentre VI risolve un problema di ottimizzazione e approssima la distribuzione a posteriori con il risultato di tale ottimizzazione. Infatti, l'Inferenza Variazionale approssima la distribuzione a posteriori in una densità più semplice, questo riduce notevolmente i tempi di esecuzione

dell'algoritmo. Il metodo MCMC consente però di campionare dalla distribuzione a posteriori esatta, e presenta soltanto una misura d'errore Monte Carlo, che, con un numero di campioni adeguato, può risultare piccolo. Il metodo VI invece approssima la forma funzionale della distribuzione a posteriori, per cui è presente una fonte d'errore che non può essere in nessun modo arginata.
Per ulteriori comparazioni tra i due approcci ci si può riferire a Blei et al. (2017).

In questo elaborato è stato utilizzato il metodo dell'ADVI (*Automatic Differentation Variational Inference*) (Kucukelbir et al., 2016). Di seguito viene brevemente introdotta l'ottimizzazione attraverso Inferenza Variazionale e nello specifico il funzionamento del metodo ADVI, implementato nella piattaforma *Stan*.

In forma generale, si consideri un modello bayesiano nel quale tutti i parametri presentano una distribuzione a posteriori. Il modello potrebbe anche avere variabili latenti oltre che parametri, per cui l'insieme di parametri e variabili latenti viene indicato con φ. I dati osservati vengono indicati con $\mathcal{Y}$. L'obiettivo dell'Inferenza Variazionale è quello di approssimare la distribuzione a posteriori $p(\varphi \mid \mathcal{Y})$. Nello specifico, si ricerca la densità $q(\varphi)$ tale da minimizzare la divergenza di Kullback-Leibler dalla distribuzione a posteriori esatta. Se non si pone alcuna restrizione su $q(\varphi)$, tale minimo si raggiunge proprio quando $q(\varphi)$ è uguale a $p(\varphi \mid \mathcal{Y})$. Si considera allora una famiglia ristretta di densità $\mathcal{L}$ e all'interno di tale famiglia si ricerca la densità $q^\star(\varphi)$ che minimizza la divergenza di Kullback-Leibler.

$$q^\star(\varphi) = \underset{q(\varphi)\in\mathcal{L}}{\operatorname{argmin}}\, KL(q(\varphi)||p(\varphi|\mathcal{Y})). \tag{2.14}$$

La distribuzione a posteriori viene quindi approssimata con $q^\star(\varphi)$. La famiglia di distribuzioni $\mathcal{L}$ deve essere tale da contenere solo distribuzioni trattabili, e allo stesso tempo deve essere abbastanza ricca da contenere distribuzioni che possono ben approssimarsi alla vera distribuzione a posteriori.
La divergenza di Kullback Leibler in equazione 2.14 equivale a

$$KL(q(\varphi)\|p(\varphi \mid \mathcal{Y})) = \mathbb{E}[\log\ q(\varphi)] - \mathbb{E}[\log\ p(\varphi \mid \mathcal{Y})], \tag{2.15}$$

Dove i valori attesi si riferiscono a $q(\varphi)$. Sostituendo l'espressione della distribuzione a posteriori

$$KL(q(\varphi)\|p(\varphi \mid \mathcal{Y})) = \mathbb{E}[\log\ q(\varphi)] - \mathbb{E}[\log\ p(\varphi, \mathcal{Y})] + \log\ p(\mathcal{Y}), \tag{2.16}$$

La distribuzione marginale $p(\mathcal{Y})$ può essere di difficile derivazione o può richiedere un elevato costo computazionale. Per tale ragione, viene massimizzata una funzione obiettivo che rappresenta il limite inferiore rispetto alla divergenza di KL, denominata *evidence lower bound* (ELBO).

$$ELBO(q) = \mathbb{E}[\log\ p(\mathcal{Y}, \varphi)] - \mathbb{E}[\log\ q(\varphi)]. \tag{2.17}$$

L'ELBO equivale alla divergenza di KL cambiata di segno con l'aggiunta del termine $\log p(\mathcal{Y})$. Massimizzare l'ELBO equivale a minimizzare la divergenza di KL.
L'equazione 2.17 può essere riscritta nel modo seguente

$$\begin{aligned} ELBO(q) = \mathbb{E}[\log\ p(\varphi)] + \mathbb{E}[\log\ p(\mathcal{Y} \mid \varphi)] - \mathbb{E}[\log\ q(\varphi))] = \\ = \mathbb{E}[\log\ p(\mathcal{Y} \mid \varphi)] - KL(q(\varphi) \| p(\varphi)). \end{aligned}$$

Questa riscrittura permette di sottolineare che la massimizzare l'ELBO equivale a massimizzare il primo termine, che porta alla scelta di distribuzioni vicine al valore atteso della verosimiglianza, e minimizzare il secondo termine, che porta a distribuzioni vicine alla distribuzione a priori. Tale funzione obiettivo presenta quindi l'usuale compromesso tra verosimiglianza e priori.

Nell'Inferenza Variazionale classica la famiglia che approssima la distribuzione viene scelta in modo da essere in accordo con la distribuzione a priori, alternativamente viene scelta *ad-hoc* rispetto al modello. Definire una procedura automatica non è perciò semplice. Il metodo che segue l'ADVI è una procedura iterativa per modelli differenziabili basata sull'integrazione Monte Carlo e sull'ottimizzazione stocastica, che riesce attraverso opportune trasformazioni delle variabili presenti entro al modello ad applicare un algoritmo di Inferenza Variazionale stocastico che procede in modo automatico. L'algoritmo ADVI opera dentro la piattaforma *Stan* in modo automatico per qualsiasi modello. Questo algoritmo utilizza un metodo di ottimizzazione stocastica e produce un'approssimazione in più rispetto ad altri algoritmi che applicano un algoritmo di Inferenza Variazionale, poiché il gradiente non viene calcolato in modo esatto, ma viene calcolato il suo valore atteso ottenuto attraverso integrazione Monte Carlo. I metodi MCMC e ADVI presentano entrambi una componente stocastica, ma il metodo ADVI riporta maggiori approssimazioni rispetto al metodo MCMC, poiché campiona da una distribuzione a posteriori approssimata e all'interno dell'algoritmo approssima anche il gradiente da calcolare. Quest'ultimo metodo ha però un elevato guadagno in termini di tempo d'esecuzione, rispetto al canonico MCMC.
Il metodo ADVI procede in primis alla trasformazione del supporto di φ in un dominio non vincolato operando una trasformazione sui parametri stessi, e la massimizzazione dell'ELBO avviene rispetto ai parametri trasformati ζ. Questa trasformazione presenta il vantaggio di poter selezionare una distribuzione variazionale q senza dover incorrere in problemi legati ai supporti vincolati. La densità approssimata scelta in questo caso per i parametri trasformati ζ è la distribuzione normale, che im-

plica una distribuzione variazionale di tipo non-normale per i parametri originali.

Si passa ora alla specificazione della famiglia di densità approssimate $\mathcal{L}$, le due principali famiglie che verranno brevemente trattate sono la *Mean-Field* e la *Full-rank*. La famiglia *Mean-Field* tratta la distribuzione congiunta delle variabili latenti come prodotto delle singole distribuzioni. In questo caso ci si è ristretti alla distribuzione normale per le variabili trasformate, per cui si avrà:

$$q(\zeta, \phi) = N(\zeta; \mu, \sigma^2) = \prod_{k=1}^{K} N(\zeta_k; \mu_k, \sigma_k^2),$$

dove $\phi = (\mu_1, \ldots, \mu_k, \sigma_1^2, \ldots, \sigma_k^2)$. Si assume quindi indipendenza tra le variabili latenti trasformate, e ognuna di esse risulta essere governata dal fattore corrispondente q_k. In fase di ottimizzazione, sono quindi le singole densità q_j ad essere ottimizzate in modo da massimizzare l'ELBO. Questo tipo di assunzione porta ad una stima appropriata della densità marginale dei parametri, tuttavia sottostima la correlazione a posteriori tra di essi, e quindi la varianza dei parametri e delle variabili latenti viene sottostimata.

La famiglia di densità approssimate *Full-rank* rappresenta una generalizzazione della famiglia *Mean-field*, nel quale si considera come distribuzione congiunta delle variabili latenti ζ trasformate la distribuzione normale multivariata

$$q(\zeta, \phi) = N(\zeta; \mu, \Sigma).$$

Questo metodo fornisce un'approssimazione a posteriori più accurata rispetto al metodo precedente, che richiede però un tempo di esecuzione e costo computazionale maggiore.

Si effettua poi un'ultima trasformazione al fine di passare dalla distribuzione normale congiunta per i parametri ζ alla distribuzione normale standard. Si definisce quindi la trasformazione $\eta = S_{\mu,\omega}(\zeta)$, ponendo $\omega = \log(\sigma)$; la trasformazione sarà diversa tra famiglia *Mean-field* e *Full-rank*. Con quest'ultima riparametrizzazione si riesce ad ottenere una formulazione dell'ELBO derivabile in modo automatico. L'equazione 2.17 viene quindi riscritta rispetto alla trasformazione delle variabili latenti identificata da η, e verrà massimizzata rispetto a μ e ω. Il valore atteso dell'ELBO riscritto in questo modo può essere calcolato empiricamente attraverso il metodo di integrazione Monte Carlo.

Dopo aver descritto tutti gli aspetti dell'inferenza variazionale e del metodo ADVI, si può procedere ad una breve trattazione dell'algoritmo applicato da ADVI per massimizzare l'ELBO.

Il meotodo ADVI utilizza al suo interno l'algoritmo dell'ascesa del gradiente, per cui l'aggiornamento delle quantità μ e ω viene calcolato come il loro valore al passo precedente sommato al gradiente calcolato nell'iterazione considerata moltiplicato per un tasso d'aggiornamento. In realtà, l'algoritmo considera il valore atteso del gradiente, che viene calcolato empiricamente via integrazione Monte

Carlo. Per questa ragione, l'algoritmo prevede la simulazione di M campioni tratti dalla distribuzione di η, che vengono poi ritrasformati per ottenere ζ, operando una trasformazione inversa rispetto alla standardizzazione. Con gli M campioni a disposizione, vengono calcolati i gradienti medi rispetto a μ e ω e avendo questi, si procede all'aggiornamento dei parametri come descritto sopra. L'algoritmo converge quando la differenza tra i valori di μ e ω sono inferiori rispetto ad una certa soglia.

2.8 Criteri di selezione dei parametri di regolazione

Tutti i modelli presentati nelle sezioni 2.4, 2.5 e 2.6 richiedono la selezione della complessità del modello in riferimento alla dimensione d della distribuzione normale multivariata associata alle posizioni latenti, e per il modello a distanze latenti con mistura anche il numero di gruppi. I criteri di selezione scelti sono anch'essi basati sull'inferenza bayesiana, e sono il criterio WAIC, *loo*, (Vehtari et al., 2016), *pseudo-BMA* (*Bayesian Model Averaging*), (Yao et al., 2018), tutti implementati entro l'omonima libreria *loo*, disponibile per la piattaforma *Stan*. Il criterio WAIC penalizza il logaritmo della densità predittiva calcolata punto per punto rispetto alla varianza a posteriori del logaritmo della densità predittiva per ogni punto, la quale converge al numero effettivo di parametri del modello. Tale criterio opera un bilanciamento tra densità predittiva e numero di parametri effettivo analogo a quello svolto dal criterio dell'AIC.

$$\begin{aligned} \widehat{elpd}_{waic} &= \widehat{lpd} - \hat{p}_{waic}, \\ \widehat{lpd} &= \sum_{t=1}^{T}\sum_{i=1}^{N_v}\sum_{j=i+1}^{N_v} \log\left(\frac{1}{B}\sum_{b=1}^{B}(\pi^b_{i,j,t})^{y_{i,j,t}}(1-\pi^b_{i,j,t})^{1-y_{i,j,t}}\right), \\ p_{waic} &= \sum_{t=1}^{T}\sum_{i=1}^{N_v}\sum_{j=i+1}^{N_v}\left(\frac{1}{B-1}\sum_{b=1}^{B}\left(y_{i,j,t}\log\frac{\pi^b_{i,j,t}}{\overline{\pi_{i,j,t}}} + (1-y_{i,j,t})\log\frac{1-\pi^b_{i,j,t}}{1-\overline{\pi_{i,j,t}}}\right)^2\right). \end{aligned} \tag{2.18}$$

Dove con $\overline{\pi_{i,j,t}}$ ci si riferisce alla media dei B elementi $\pi_{i,j,t}$ campionati, con *elpd* si intende il logaritmo della densità predittiva attesa per ogni punto di un nuovo dataset, che richiederebbe di conoscere la distribuzione del vero processo generatore dei dati. Questa quantità viene allora stimata in modo approssimato dal criterio WAIC come in Equazione 2.18. Con *lpd* ci si riferisce al logaritmo della densità predittiva per ogni punto calcolata empiricamente, usando quindi delle simulazioni da tale distribuzione.

Il criterio del *loo* opera una stima puntuale del logaritmo della densità predittiva a posteriori

tramite convalida incrociata *leave-one-out*.

$$elpd_{loo} = \sum_{t=1}^{T} \sum_{i=1}^{N_v} \sum_{j=i+1}^{N_v} \log p(y_{i,j,t} \mid y_{-i,j,t}),$$

dove

$$p(y_{i,j,t} \mid y_{-i,j,t}) = \int p(y_{i,j,t} \mid \pi_{i,j,t}) p(\pi_{i,j,t} \mid y_{-i,j,t}) d\pi_{i,j,t},$$

indica la densità predittiva calcolata attraverso i dati, senza l'i-esima osservazione, (in questo contesto il dataset è stato ristretto alla matrice triangolare superiore per ogni tempo, cosicché l'eliminazione di una riga dal dataset portasse al risultato voluto) e la stessa densità predittiva considerandola invece entro il dataset. Nella pratica, la quantità $elpd_{loo}$ viene quindi calcolata come

$$\widehat{elpd}_{loo} = \sum_{t=1}^{T} \sum_{i=1}^{N_v} \sum_{j=i+1}^{N_v} \log \left(\frac{1}{B_i} \sum_{b_i=1}^{B_i} (y_{i,j,t}) \log \pi_{i,j,t}^{b_i} + (1 - y_{i,j,t}) \log(1 - \pi_{i,j,t}^{b_i}) \right), \tag{2.19}$$

dove con B_i ci si riferisce ai campioni tratti dalla distribuzione a posteriori calcolata avendo eliminato la riga i-esima della matrice di adiacenza per tutti gli istanti temporali considerati.

Il metodo del *pseudo-Bayesian Model Averaging* consente di selezionare un modello tra K possibili assegnando un valore ad ogni modello, questo valore viene calcolato come media aritmetica,

$$w_k = \frac{\exp(\widehat{elpd}_{loo}^{k})}{\sum_{k=1}^{K} \exp(\widehat{elpd}_{loo}^{k})}, \quad k = 1, \ldots, K,$$

nel quale $\widehat{elpd}_{loo}^{k}$ indica la stima del logaritmo della densità predittiva calcolata in corrispondenza del modello k-esimo, come in Equazione 2.19. Si selezionerà il modello che riporta un valore maggiore di w_k.

Per tutte le analisi, queste metriche sono state in accordo tra di loro. Nel caso di risultati uguali per due valori della dimensione della variabile stessa è stato selezionato quello inferiore, nel rispetto del criterio di parsimonia del modello.

Capitolo 3

Simulazioni

I modelli descritti nel capitolo 2 possono essere testati attraverso alcuni studi di simulazione, situazione in cui si conosce il processo generatore dei dati. Sono stati effettuati due studi di simulazione che si avvicinassero alle assunzioni sulle quali sono basati i modelli appena considerati, in modo da poter confrontare il loro andamento sia quando il processo generatore dei dati è vicino a quello usato per modellare la rete, sia quando questo non avviene. Per entrambe le simulazioni, sono stati considerati 80 nodi e 14 istanti temporali. La rete dinamica è stata simulata a partire dalla simulazione delle coordinate latenti e dalla media del processo (rispetto al logaritmo della quota di ogni $\pi_{i,j,t}$), a partire da assunzioni distributive che verranno specificate per ogni scenario. Dopo aver simulato queste quantità, è stato calcolato il predittore lineare seguendo una formulazione tra quelle in equazione 2.9 o equazione 2.5. Infine, la rete dinamica è stata calcolata sulla base del predittore lineare (applicando prima la trasformazione inversa al logit, per riportare tale valore al dominio $(0, 1)$), in modo da avere un arco tra un nodo i e un nodo j solo se il predittore lineare corrispondente fosse maggiore del valore soglia 0.5.

Vengono di seguito presentate le assunzioni distributive sotto le quali sono state simulate le due reti dinamiche.

- Nel primo scenario le coordinate latenti sono state simulate usando la stessa struttura del modello a distanze latenti dinamico (sezione 2.5), ponendo $d = 3$ e considerando come media per le posizioni latenti al tempo iniziale tre valori distinti, pari rispettivamente a vettori d-dimensionali aventi componenti tutte pari a 0 per i primi 20 nodi, 0.5 per i successivi 30 e 0.9 per i rimanenti 30, nel tentativo di creare tre gruppi entro la rete, mentre il termine di intercetta è stato generato rispettando ancora una struttura autoregressiva del primo ordine, distanziandosi invece

dall'assunzione effettuata in sezione 2.5. Questa scelta è motivata dal fatto che si vuole indagare la capacità predittiva dei modelli non rispettando esattamente l'assunzione distributiva per nessuno di essi. Per le coordinate latenti è stato posto σ^2 e τ^2 pari a 0.5, mentre per il termine di intercetta è stata usata una varianza pari a 0.25, ponendo come valore iniziale 0.2. Il parametro γ è stato scelto pari a 0.8. Dopo aver generato in questo modo le posizioni latenti e l'intercetta per tutti i tempi, è stato calcolato il predittore lineare come in Equazione 2.9, non considerando però il termine relativo alle variabili esplicative. Con questa formulazione il grado di separabilità tra i tre gruppi non risulta essere particolarmente accentuato.

- Nel secondo scenario le coordinate latenti sono state simulate sulla base della struttura del modello a proiezioni latenti dinamico (sezione 2.4), ponendo $d = 4$ e differenziando ancora 3 gruppi, con le stesse suddivisioni usate nel primo scenario di simulazione, ponendo media diversa delle coordinate latenti e imponendo una struttura di dipendenza temporale diversa. Il termine di intercetta è stato generato rispettando ancora una struttura autoregressiva del primo ordine, non seguendo quindi la distribuzione originale proposta in sezione 2.4. Per le coordinate latenti sono state usate tre medie distinte, pari a vettori T-dimensionali con componenti tutte rispettivamente pari a 0.5 per il primo gruppo, 2.5 per il secondo, -1 per il terzo, mentre i parametri di scala per la struttura della covarianza sono stati posti rispettivamente pari a $0.3, 0.5, 0.8$. Il termine di intercetta è stato fissato al tempo 1 pari a 0.3 ed è stato generato tramite una distribuzione normale multivariata con varianza 0.25. Il predittore lineare è stato calcolato seguendo la formula riportata in Equazione 2.5 (senza considerare il termine legato alle variabili esplicative). La rete dinamica simulata presenta una struttura diversificata a seconda dei tre gruppi: il primo presenta un livello di assortatività medio, la seconda ha una struttura totalmente assortativa e presenta una comunità totalemente interconnessa, l'ultimo gruppo ha un buon livello di assortatività. Il numero di connessioni tra gruppi è invece ridotto. Si vuole verificare se in questo caso il modello con mistura sarà in grado di identificare correttamente questi gruppi e riprodurre queste strutture di connessione entro le comunità.

Inoltre, avendo usato due processi generatori delle coordinate latenti nel primo caso propri del metodo a distanze latenti dinamico, nel secondo caso del modello a proiezioni latenti dinamico, si vuole verificare se i modelli nella pratica siano in grado di avere delle buone prestazioni anche quando il processo generatore dei dati non è quello assunto nel modello.

3.1 Approccio all'analisi

L'implementazione dei modelli presentati nel capitolo precedente è stata effettuata tramite l'ausilio della piattaforma *Stan*. Per svolgere un'analisi dei risultati ottenuti per i tre modelli implementati è stata ricavata la distribuzione predittiva a posteriori per delle nuove osservazioni, nel presente contesto i campioni tratti da questa distribuzione corrispondono a delle reti. Nello specifico, sono state simulate $B = 1000$ reti dinamiche per i 14 tempi considerati dalla densità predittiva a posteriori. La bontà d'adattamento del modello è stata misurata in due modi:

- sono state confrontate le statistiche a livello di nodo e di rete con le stesse misurate nella rete osservata;

- le reti simulate sono state confrontate con la rete realmente osservata attraverso la matrice di confusione.

Per il primo approccio, le statistiche di rete e di nodo possono essere calcolate rispetto ai campioni tratti dalla densità predittiva a posteriori, quindi nelle reti simulate. In questo modo, la distribuzione della statistica calcolata rispetto alla densità predittiva a posteriori può essere ricavata in modo approssimato e può essere confrontata con il valore della statistica calcolata rispetto alla rete osservata. Si disporrà quindi di un solo valore della statistica nella rete osservata, e di una distribuzione di statistiche calcolate sulle reti simulate. Avendo una distribuzione e un valore osservato a confronto, si dovrà verificare se la distribuzione è coerente con il valore realmente osservato. Se questa condizione è verificata, il modello presenta un buon adattamento alla rete osservata e riesce a emularne le principali caratteristiche.

Per il secondo approccio proposto le reti simulate dalla densità predittiva a posteriori e la rete dinamica osservata sono state confrontate attraverso le usuali metriche di classificazione, applicando per ognuna la media rispetto ai valori ottenuti per ogni rete simulata. Nello specifico sono state riportate il tasso di errata classificazione, il tasso di falsi negativi, il tasso di falsi positivi e la metrica F1, per ognuno di essi è stata riportata non solo la media ma anche la deviazione standard. Attraverso questi elementi è possibile confrontare la bontà d'adattamento dei modelli considerati in termini di corretta identificazione degli archi e scegliere il modello che presenta una migliore rispondenza rispetto alla struttura della rete dinamica osservata.

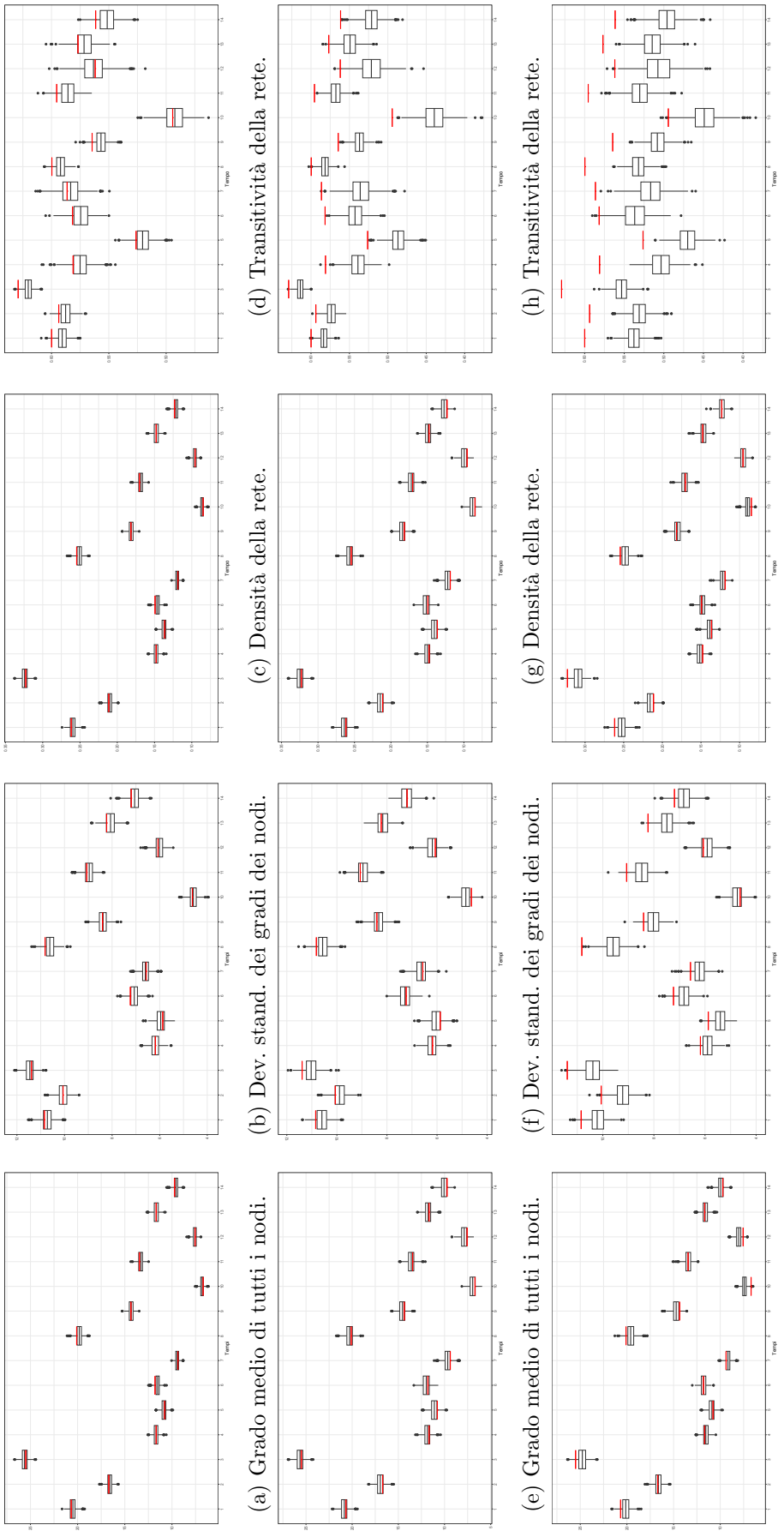
3.2 Primo scenario di simulazione

Dopo aver ottenuto la rete simulata, è stata selezionata la dimensione d per le posizioni latenti attraverso i criteri d'informazione riportati in sezione 2.8, che per modello a distanze latenti hanno correttamente selezionato $d = 3$, per il modello a proiezioni latenti invece è stato selezionato $d = 5$, infine per il modello a distanze latenti con mistura invece si seleziona $d = 3$ e $G = 1$, non è quindi stato in grado di individuare i tre gruppi creati in fase di simulazione.

I risultati in termini di capacità previsiva dei tre modelli (Tabella 3.1) evidenziano che il modello a distanze latenti risulta essere quello con tasso di errata classificazione inferiore rispetto a tutti gli altri, come ci si attendeva dal fatto che è quello che corrisponde maggiormente al processo generatore dei dati. In ogni caso, anche il modello a proiezioni latenti presenta un'ottima bontà d'adattamento, presentando solo un valore leggermente maggiore in corrispondenza del tasso di falsi negativi. Questi due andamenti sono quindi del tutto comparabili. Per quanto riguarda invece il modello a distanze latenti con mistura, si ottiene un tasso di errata classificazione doppio rispetto agli altri due modelli, ma in ogni caso si tratta di un valore abbastanza ridotto; il tasso di falsi negativi risulta essere molto maggiore rispetto a quello ottenuto negli altri due casi, quindi il modello sottostima in modo maggiore il numero di archi tra i vertici, riportando in modo coerente un valore inferiore della metrica F1. Per cercare riscontro rispetto a queste affermazioni, è possibile considerare i grafici mostrati in Figura 3.1, dal quale si evince che i primi due modelli sono coerenti rispetto ai valori realmente osservati nella rete dinamica, mentre il modello a distanze latenti con mistura presenta distribuzioni non sempre coerenti con il valore osservato della statistica. Inoltre, i primi due modelli mostrano delle distribuzioni predittive a posteriori caratterizzate da minore variabilità rispetto al modello con mistura. Per la deviazione standard dei gradi tutti e tre i modelli presentano distribuzioni molto variabili, ma mentre i primi due hanno un andamento coerente con il valore realmente osservato, il terzo sottostima anche questa quantità. Il livello di transitività non viene identificato in modo esatto da nessun modello ma in particolar modo dal terzo, il primo riesce però ad avere distribuzioni abbastanza prossime rispetto al valore realmente osservato.

	errata classificazione	*falsi positivi*	*falsi negativi*	*metrica F1*
Proiezioni latenti	0.052(0.001)	0.034(0.001)	0.129(0.002)	0.865(0.002)
Distanze latenti	0.040(0.001)	0.024(0.001)	0.106(0.002)	0.896(0.002)
Distanze latenti con mistura	0.104(0.001)	0.062(0.001)	0.306(0.005)	0.695(0.004)

Tabella 3.1: Tassi di classificazione ottenuti per il primo scenario di simulazione.



(a) Grado medio di tutti i nodi. (b) Dev. stand. dei gradi dei nodi. (c) Densità della rete. (d) Transitività della rete.

(e) Grado medio di tutti i nodi. (f) Dev. stand. dei gradi dei nodi. (g) Densità della rete. (h) Transitività della rete.

(i) Grado medio di tutti i nodi. (j) Dev. stand. dei gradi dei nodi. (k) Densità della rete. (l) Transitività della rete.

Figura 3.1: Statistiche descrittive di rete e di nodo calcolate sulle reti simulate dalla densità predittiva a posteriori. Le quattro figure riportate nella prima riga sono relative al modello a distanze latenti dinamico, quelle nella seconda riga al modello a proiezioni latenti dinamico, quelle nella terza al modello a distanze latenti con mistura.

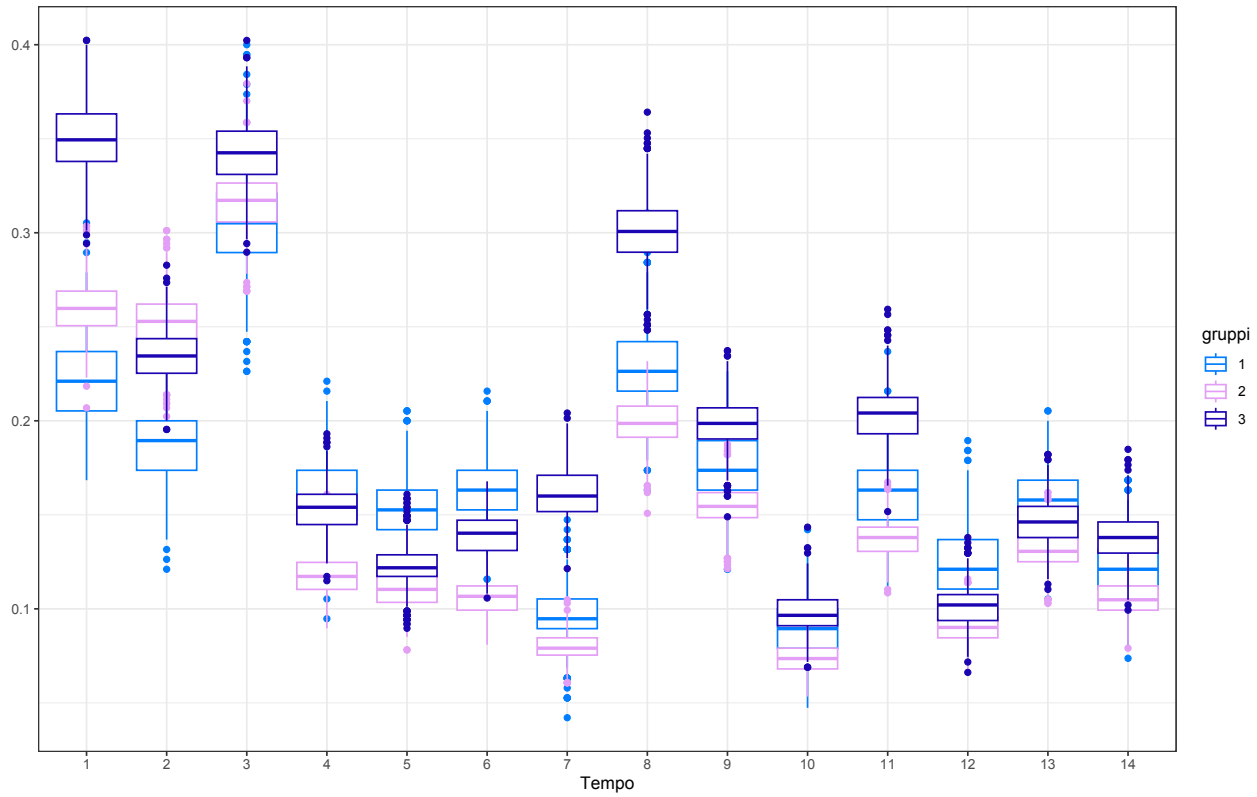


Figura 3.2: Densità media delle comunità create a priori in fase di simulazione per i tempi considerati.

Il modello a distanze latenti con mistura non ha identificato correttamente i gruppi creati a priori. Si vogliono ora mostrare le distribuzioni delle densità media delle comunità formate, per evidenziare quanto effettivamente fossero diverse (Figura 3.2). Si può in effetti notare come i tre gruppi mostrino un andamento diversificato per alcuni istanti temporali, mentre per altri istanti temporali i gruppi sono quasi sovrapposti, o comunque potrebbero essere considerati un gruppo unico che presenta elevata variabilità. Inoltre, poiché la media (almeno iniziale) delle coordinate latenti era effettivamente vicina, i gruppi non presentano un comportamento distintivo che può definirsi proprio del gruppo, come un inferiore numero di connessioni per tutti i tempi, o un gruppo che spicca per numero di connessioni in tutto il periodo di osservazione.

3.3 Secondo scenario di simulazione

Dopo aver simulato la rete dinamica, sono stati applicati i criteri di selezione esposti in sezione 2.8 per stabilire il numero di coordinate latenti, e per il terzo modello anche il numero di gruppi. Per il modello a proiezioni latenti e per quello a distanze latenti è stato selezionato $d = 4$, mentre per il modello con mistura è stato scelto $d = 2, G = 3$.

I tassi di classificazione ottenuti per i tre modelli (Tabella 3.2) sono soddisfacenti, tutti i modelli riportano un tasso di errata classificazione ridotto, i primi due a posizioni latenti tempo-dipendenti sembrano equivalersi in termini di andamento, mentre il modello con mistura ha un tasso di falsi positivi doppio rispetto agli altri due. Le metriche F1 riportate assumono tutte un valore elevato. Le distribuzioni predittive a posteriori calcolate rispetto alle statistiche di nodo e di rete (Figura 3.3) mostrano un andamento coerente tra reti simulate e valore realmente osservato in termini di densità media, media dei gradi per il modello a proiezioni latenti e a distanze latenti, anche se le distribuzioni presentano comunque molta variabilità. Il livello di transitività non viene invece correttamente identificato. Le deviazioni standard dei gradi entro le reti, per ogni tempo fissato, vengono sovrastimate dal modello a distanze latenti, sottostimate da quello a proiezioni latenti. Osservando invece le stesse distribuzioni per il modello con mistura, si vede come il modello non riesca a cogliere correttamente le caratteristiche della rete, poiché tutte le distribuzioni risultano essere non coerenti rispetto al valore realmente osservato. Le distribuzioni corrispondenti al livello di transitività sono analoghe agli altri due modelli. Sebbene i tassi di classificazione mostrino una buona prestazione di questo modello, dai grafici nel terzo pannello si vede come il modello non riesca a emulare correttamente l'evoluzione temporale delle reti. Nonostante questo, si può comunque provare ad analizzare la suddivisione in gruppi operata dal modello.

La tabella di corretta classificazione, riportata in Tabella 3.3, mostra come le classi risultano ben separabili, infatti è possibile assegnare il gruppo 1 al gruppo con media assortatività, il gruppo 2 a quello con buona assortatività, il gruppo 3 alla comunità totalmente interconnessa. Il modello stima le varianze a posteriori per i tre gruppi ponendo 0.18 come stima puntuale per il primo gruppo, in

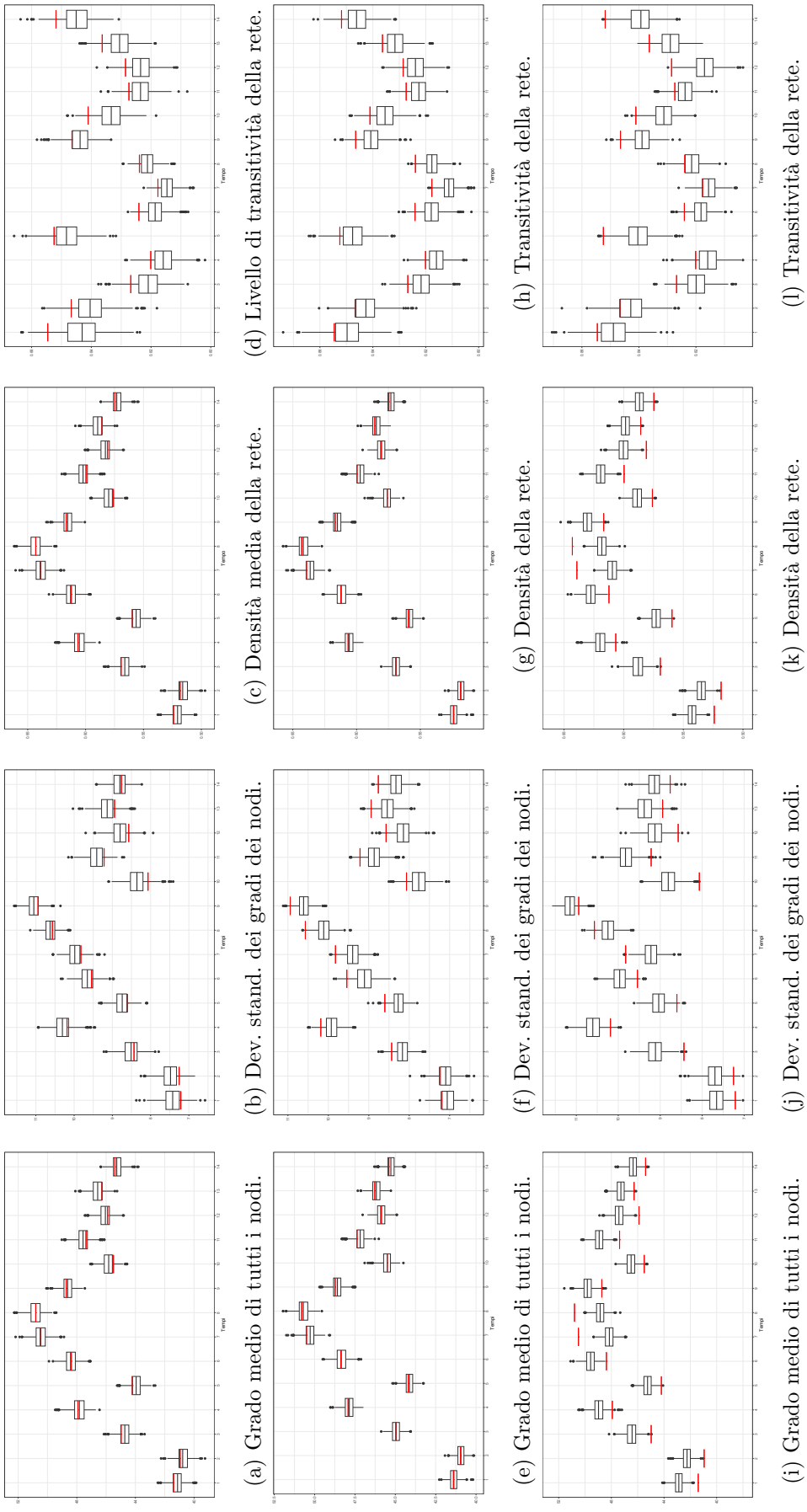
	errata classificazione	*falsi positivi*	*falsi negativi*	*metrica F1*
Proiezioni latenti	0.040(0.001)	0.046(0.002)	0.036(0.002)	0.965(0.001)
Distanze latenti	0.043(0.001)	0.051(0.002)	0.037(0.001)	0.963(0.001)
Distanze latenti con mistura	0.075(0.001)	0.101(0.002)	0.057(0.001)	0.935(0.001)

Tabella 3.2: Tassi di classificazione ottenuti per il secondo scenario di simulazione

accordo col fatto che si tratta del gruppo con più unità classificate erroneamente, 0.028 per il secondo gruppo e 0.02 per il terzo, con stime quindi abbastanza ridotte; le medie stimate a posteriori sono pari a $-0.706, 0.093$ per il primo gruppo, $0.714, -0.036$ per il secondo, $-0.066, -0.016$ per il terzo. Si verifica ora se il modello è stato in grado di cogliere la struttura assortativa impartita alla rete. In Figura 3.4 è possibile confrontare le densità medie entro le tre comunità create a priori (pannello superiore) con quelle identificate dal modello (pannello inferiore), quindi rispettando le assegnazioni mostrate in Tabella 3.3. Come si può vedere, le comunità a priori presentano tre livelli distinti di connessioni entro le stesse, ma comunque possiedono tutte una struttura fortemente assortativa. Il modello riesce ad identificare correttamente i gruppi 2 e 3, riuscendo conseguentemente a replicare in modo soddisfacente il numero di connessioni entro queste comunità; il primo gruppo è quello che presenta più unità classificate in modo non corretto, e avendo le comunità poche connessioni tra gruppi diversi, la densità media entro questo gruppo risente di questa caratteristica, portando ad una sua sottostima. Nonostante ciò, la classificazione porta ad un risultato soddisfacente. Rispetto alla simulazione precedente, i livelli di densità dei tre gruppi creati a priori mantengono la stessa gerarchia per tutti i tempi, e presentano caratteristiche più marcate. C'è quindi un livello di separabilità più apprezzabile rispetto al primo scenario. I gruppi identificati dal modello invece non rispettano sempre questa gerarchia, come si può vedere in corrispondenza dei tempi 6,7,8, ma colgono comunque le principali caratteristiche che contraddistinguono le comunità.

	1	2	3
Media assortatività	15	0	5
Buona assortatività	10	20	0
Completamente connessa	2	0	28

Tabella 3.3: Assegnazioni ai gruppi effettuate dal modello a distanze latenti con mistura per il secondo scenario di simulazione.

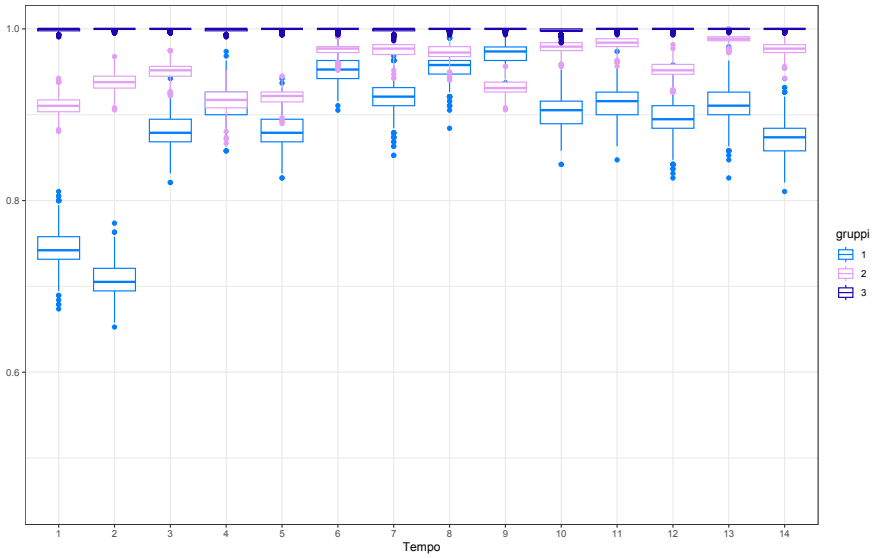


(a) Grado medio di tutti i nodi. (b) Dev. stand. dei gradi dei nodi. (c) Densità media della rete. (d) Livello di transitività della rete.

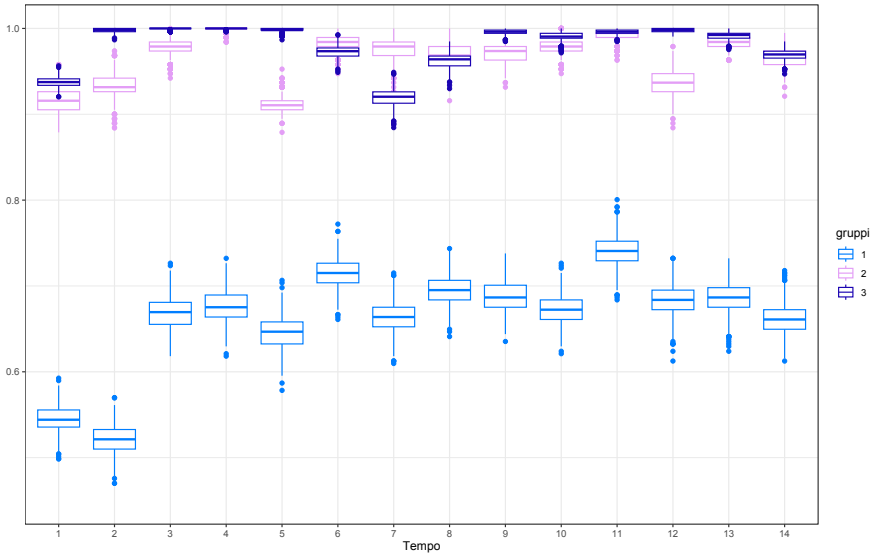
(e) Grado medio di tutti i nodi. (f) Dev. stand. dei gradi dei nodi. (g) Densità della rete. (h) Transitività della rete.

(i) Grado medio di tutti i nodi. (j) Dev. stand. dei gradi dei nodi. (k) Densità della rete. (l) Transitività della rete.

Figura 3.3: Statistiche descrittive di rete e di nodo calcolate sulle reti simulate dalla densità predittiva a posteriori. Le quattro figure riportate nella prima riga sono relative al modello a distanze latenti dinamico, quelle nella seconda riga al modello a proiezioni latenti dinamico, quelle nella terza al modello a distanze latenti con mistura.



(a) Andamento della densità media dei tre gruppi imposto alla rete a priori.



(b) Andamento della densità media dei tre gruppi identificati dal modello.

Figura 3.4: Densità predittiva a posteriori della densità media calcolata rispetto alle tre comunità ad ogni istante temporale. Nel pannello superiore si mostra l'andamento ottenuto in corrispondenza delle tre comunità vere imposte, nel pannello inferiore l'andamento rispetto ai tre gruppi identificati dal modello.

3.4 Commenti finali

Le prestazioni riscontrate nei modelli per i due scenari di simulazione presentati sono soddisfacenti per quanto riguarda il modello a proiezioni latenti dinamico e il modello a distanze latenti dinamico, poiché non solo riescono in entrambi i casi a presentare un errore di classificazione molto ridotto e a riproporre correttamente le proprietà della rete, ma le distribuzioni presentano anche poca variabilità. Il modello con mistura ha riportato delle buone prestazioni in termini di classificazione degli archi, ma in termini di densità predittive a posteriori per le statistiche descrittive considerate non ha risultati soddisfacenti come gli altri due modelli. Inoltre, il modello non ha identificato i gruppi imposti nel primo scenario di simulazione, come ci si aspettava anche in fase di simulazione delle reti. Questo risultato è certamente da attribuirsi al piccolo livello di separabilità imposto e non dal processo generatore delle coordinate latenti, poiché i risultati ottenuti per i due modelli dinamici sono molto simili in entrambi i casi, per cui simulando in un modo o nell'altro non si notano differenze sostanziali in termini di comportamento delle reti.

Quando il livello di separazione tra i gruppi è adeguato, il modello con mistura è in grado di identificare con un buon margine le giuste assegnazioni ai gruppi, producendo un piccolo errore di classificazione.

Inoltre, da queste simulazioni si può evincere che tutti i modelli utilizzati non sono in grado di identificare correttamente il livello di transitività presente nel modello, poiché le densità predittive a posteriori per questa statistica risultano molto variabili e comunque non sempre coerenti con il valore misurato nella rete reale. Ci si aspetta quindi di ritrovare gli stessi risultati anche nell'applicazione al dataset reale.

Capitolo 4

Applicazione al dataset *Social Evolution*

Nel presente capitolo verranno mostrati i risultati delle analisi eseguite sul dataset relativo a degli studenti del MIT presentato nell'articolo di Madan et al. (2011). I tre modelli che verranno adattati alla rete sono nuovamente il modello a proiezioni latenti dinamico, il modello a distanze latenti dinamico e il modello a distanze latenti con mistura. L'approccio all'analisi sarà quello descritto in sezione 3.1. In questo caso però, oltre a svolgere un'analisi rispetto alla bontà d'adattamento dei modelli rispetto agli stessi dati usati per adattare il modello, si valuteranno le prestazioni rispetto alla previsione della rete futura. A seguito dell'analisi esplorativa mostrata nella sezione 1.4, è stato ritenuto ragionevole restringere il numero di reti statiche da analizzare fino al tempo 32, ed effettuare una previsione per il tempo 33 (*one step ahead)*. É importante sottolineare che, in fase di analisi, nel predittore lineare per la rete Y_t è stata inserita l'informazione relativa alle variabili esplicative tempo-dipendenti al tempo $t-1$, per non incorrere in problemi legati a variabili *leaker* per la previsione. La previsione un passo in avanti è quindi legata anche a questa scelta.

Per valutare la bontà predittiva dei tre modelli si procederà in modo analogo a quanto proposto nel secondo approccio presentato in sezione 3.1, simulando ancora $B^* = 1000$ reti per il tempo 33 e riportando le metriche di classificazione. Nella previsione di una nuova rete, si ritiene che un modello che prevede un numero inferiore di archi rispetto a quelli realmente presenti, ma che classifica correttamente quelli previsti, sia migliore di un modello che invece commette più errori in termini di falsi positivi, poiché per la rete sociale dinamica considerata si ritiene meno accettabile predire archi che in realtà non saranno presenti, rispetto al caso di predirne meno ma in modo corretto.

Un'ulteriore analisi che verrà di seguito proposta è il confronto tra i modelli presentati nel capitolo 2, che definiscono un *array* a tre dimensioni per le posizioni latenti e assumono correlazione temporale tra di esse, e gli stessi modelli con posizioni latenti costanti nel tempo. Vale la pena considerare un

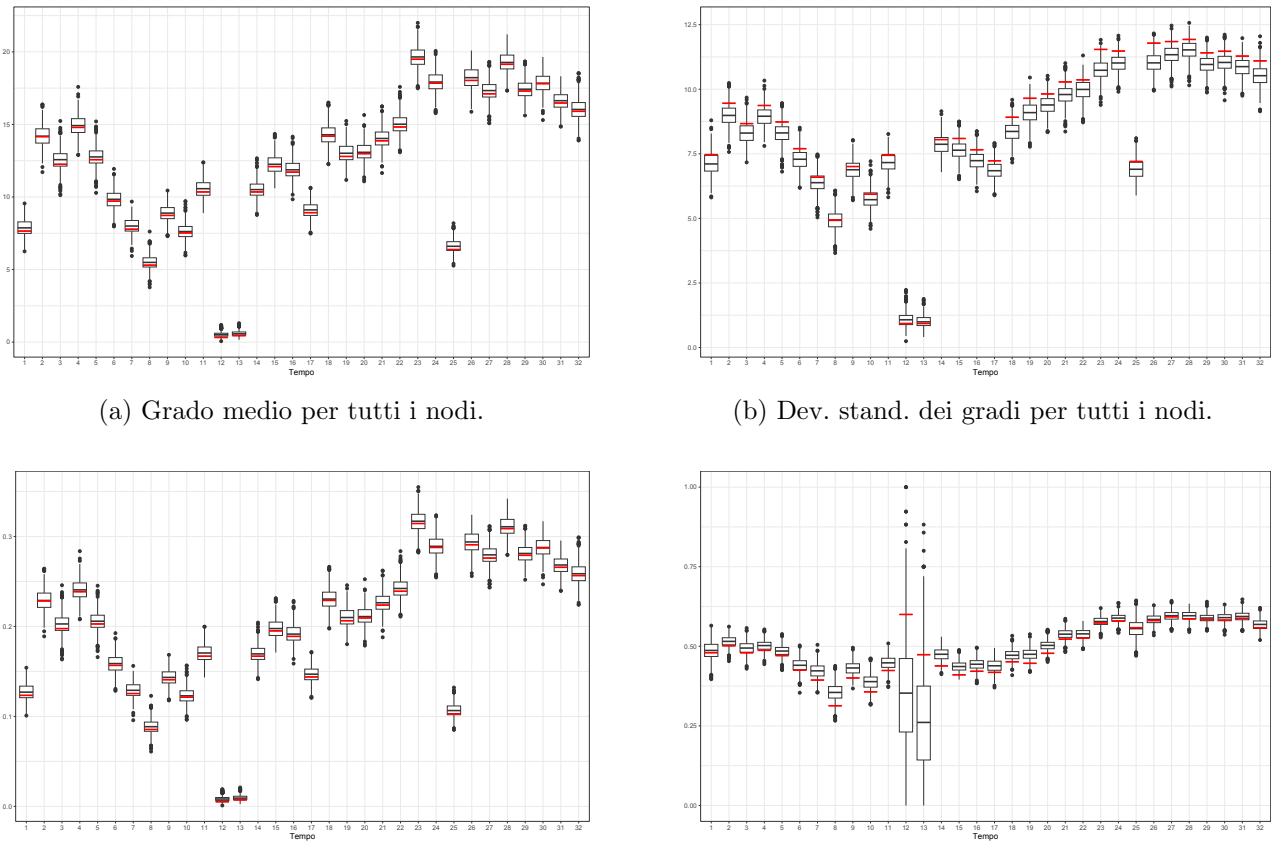
	errata classificazione	*falsi positivi*	*falsi negativi*	*metrica F1*
$\mathbf{x}_t$ variabili, senza Z_t	0.119(0.001)	0.074(0.002)	0.290(0.004)	0.700(0.003)
$\mathbf{x}_t$ variabili, con Z_t	0.107(0.001)	0.068(0.002)	0.271(0.005)	0.725(0.003)
$\mathbf{x}_t$ costanti, senza Z_t	0.176(0.001)	0.118(0.002)	0.410(0.006)	0.551(0.003)
$\mathbf{x}_t$ costanti, con Z_t	0.175(0.001)	0.117(0.002)	0.406(0.006)	0.561(0.003)

Tabella 4.1: Tassi di classificazione del modello a proiezioni latenti dinamico.

modello ridotto di questo tipo poiché la complessità del modello, dato dal numero di parametri coinvolti sarebbe notevolmente ridotto con questa assunzione semplificatrice. Come è chiaro pensare, il modello ridotto avrà un andamento peggiore, dando meno flessibilità all'evoluzione delle probabilità di connessione negli istanti temporali, si vuole però indagare di quanto l'adattamento peggiori in termini prima del secondo approccio proposto in sezione 3.1, (metriche di classificazione) ed eventualmente del primo, (distribuzione empirica delle statistiche di nodo) rispetto al metodo proposto nel capitolo 2. Questo consente di mettere in luce quali siano gli effettivi punti di forza dei modelli a variabili tempo-dipendenti, ovvero quali elementi siano meglio rappresentati grazie a tale flessibilità.

4.1 Applicazione: Modello a proiezioni latenti dinamico

Il primo modello che verrà esaminato è quello a proiezioni latenti dinamico, sviluppato da Durante & Dunson (2014). Per verificare l'importanza sia delle variabili esplicative sia della flessibilità data dall'uso di coordinate latenti tempo-dipendenti, di seguito queste tre possibilità verranno confrontate in termini di bontà d'adattamento e capacità previsiva, valutate rispetto ai dati usati per adattare il modello. Le distribuzioni a priori sono state scelte di tipo non informativo, per il parametro τ_h è stata usata una distribuzione gamma inversa con iperparametri $a = 2$, $b = 1$, che forniscono al parametro una distribuzione con varianza non finita. Sono state comunque eseguite alcune prove per verificare quanto gli iperparametri a e b influenzassero i risultati, ma si è giunti sempre agli stessi risultati a posteriori. Per i due modelli con coordinate latenti tempo-dipendenti, con e senza variabili esplicative, è stato selezionato un numero di coordinate latenti pari a $d = 7$. Le distribuzioni a priori per i coefficienti β considerati prevedevano media a priori pari a 0.5, poiché ci si aspetta un effetto positivo rispetto alle variabili considerate, e una varianza pari a 1000, per cui anche queste sono state scelte di tipo non informativo. Aumentare la varianza infatti non ha avuto alcun tipo di effetto rispetto ai risultati ottenuti della distribuzione a posteriori. Infine, per il modello con posizioni latenti costanti,



(a) Grado medio per tutti i nodi.

(b) Dev. stand. dei gradi per tutti i nodi.

(c) Densità della rete.

(d) Transitività della rete.

Figura 4.1: Statistiche di nodo e di rete misurate per il modello di proiezione dinamico con posizioni latenti variabili nel tempo con variabili esplicative. Le linee rosse rappresentano il valore delle statistiche realmente osservato nella rete.

con e senza variabili esplicative, è stato selezionato $d = 15$. La distribuzioni a priori per le posizioni latenti prevede una varianza a priori pari a 5 per ogni componente $\mathbf{x}_{i,h}$, mentre gli altri iperparametri selezionati sono gli stessi usate per i due modelli con coordinate latenti variabili nel tempo.

In Tabella 4.1 vengono mostrati i risultati in termini di bontà predittiva dei quattro modelli considerati. Si può vedere come i tassi di errata classificazione si mantengano ridotti per tutti e quattro i modelli, questo porta a dire che il modello abbia in generale un buon adattamento rispetto al problema affrontato. Si possono ora confrontare i tassi ottenuti tra i due macro modelli con posizioni latenti costanti e variabili, che verranno rispettivamente chiamati di seguito come 'ridotto' e 'completo'. Si può innanzitutto osservare che il tasso dei falsi negativi aumenta in modo importante per il modello ridotto, ed anche il tasso dei falsi positivi aumenta. Questo elemento si ripercuote nella metrica F1, che si riduce drasticamente per il modello ridotto. Questa metrica è infatti molto importante in questo contesto poiché viene calcolata come trasformata della precisione e del recupero, perciò pesa maggiormente il numero di veri positivi rispetto ai veri negativi. Si può inoltre vedere che le variabili

	Media	Deviazione Standard	Inferiore 0.05	Superiore 0.95
Chiamate effettuate	0.616	0.075	0.492	0.741
Messaggi scambiati	0.643	0.151	0.399	0.897
Anno accademico frequentato: GRT	-0.139	0.035	-0.196	-0.084
Anno accademico frequentato: Junior	-0.637	0.038	-0.700	-0.579
Anno accademico frequentato: Senior	-0.771	0.034	-0.828	-0.717
Anno accademico frequentato: Secondo anno	-0.158	0.029	-0.206	-0.110
Anno accademico frequentato: Sconosciuto	0.113	0.153	-0.128	0.374
Piano del dormitorio: 289.2	-1.070	0.040	-1.140	-1.000
Piano del dormitorio: 289.3	-1.210	0.033	1.270	-1.150
Piano del dormitorio: 289.4	-1.690	0.039	-1.750	-1.620
Piano del dormitorio: 290.2	-1.070	0.054	-1.160	-0.981
Piano del dormitorio: 290.3	-0.844	0.036	-0.901	-0.783
Piano del dormitorio: 290.4	-1.840	0.061	-1.950	-1.740
Piano del dormitorio: Sconosciuto	-1.360	0.066	-1.470	-1.250
Residenza nello stesso piano	1.160	0.034	1.100	1.210
Stesso anno accademico frequentato	-0.383	0.035	-0.441	-0.322

Tabella 4.2: Coefficienti delle variabili esplicative per il modello a proiezioni latenti dinamico con posizioni latenti tempo-dipendenti.

esplicative, per entrambi i macro modelli, portano ad un lieve miglioramento in termini di capacità previsiva della rete dinamica. A fronte dei tassi di classificazione ottenuti, il modello che risulta preferibile è quello completo con variabili esplicative.

Le distribuzioni della densità predittiva a posteriori per il grado medio, deviazione standard dei gradi, densità media, transitività media verranno mostrate solo per questo modello. Questi grafici sono stati riportati in Figura 4.1. Il modello riesce a emulare correttamente le proprietà della rete dinamica osservata, infatti le tre distribuzioni in Figura 4.1a, Figura 4.1c e 4.1d sono centrate rispetto al valore rilevato nella rete dinamica osservata, mentre si osserva una sottostima della deviazione standard dei gradi dei nodi ad ogni tempo (Figura 4.1b). Si nota inoltre che i grafici relativi alla media del nodo e alla densità presentano distribuzioni con un certo livello di variabilità, sopratutto per i tempi a partire dal numero 23. Anche riferendosi agli stessi grafici ottenuti in sede di simulazione, le distribuzioni predittive a posteriori sono molto più variabili. Ricordando che in Tabella 4.1 il numero di connessioni della rete veniva sovrastimato, questa considerazione è compatibile anche con la sottostima della deviazione standard dei gradi, che infatti nelle reti simulate restano più costanti, mentre nella rete

osservata è presente più variabilità. Anche alla luce di queste densità predittive a posteriori si può affermare che il modello presenti un buon adattamento rispetto alla rete dinamica in esame.

Vengono ora presentati i coefficienti ottenuti rispetto alle variabili esplicative introdotte nel modello (Tabella 4.2). Si può subito notare che la maggior parte dei coefficienti riporta intervalli di credibilità a posteriori *equi-tailed* che non comprendono lo zero al livello 0.05. I coefficienti riportati in Tabella 4.2 esprimono l'effetto sul logaritmo della quota di un incremento unitario della variabile esplicativa, fermo restando le altre variabili. I coefficienti positivi hanno un effetto positivo sulla probabilità di connessione, infatti per esempio si può affermare che la quota di probabilità di avere una connessione tra due nodi che si sono scambiati almeno una telefonata entro le due settimane è $\exp(0.616) = 1.85$ volte la stessa quota per i soggetti che non si sono scambiati nessuna telefonata, fermo restando le altre variabili esplicative. I coefficienti relativi all'anno accademico frequentato consentono di affermare che le matricole presentano un numero di connessioni maggiore rispetto a tutte le altre classi. Infatti, l'effetto (sempre negativo) più vicino allo zero si ha per i ragazzi del secondo anno (*Sophomore*), per i quali la quota di probabilità di aver incontrato un altro soggetto è di $\exp(-0.158) = 0.853$ volte la stessa per le matricole. Se invece ci si concentra sul piano di residenza, si può ancora affermare che la modalità di riferimento (289.1) sembra avere un numero di connessioni maggiore. Il coefficiente relativo alla convivenza nello stesso piano ha invece un effetto positivo.

Si prosegue ora considerando la bontà d'adattamento dei modelli in termini di previsione di una nuova rete, che verrà considerata solo per i modelli con variabili latenti tempo-dipendenti, a fronte di quanto emerso nella sezione precedente, con e senza variabili esplicative, sempre sotto le stesse distribuzioni esposte in precedenza. Questi due modelli verranno ancora confrontati per verificare se la presenza di variabili esplicative possa invece essere importante in termini di capacità di previsione di una nuova rete.

In Tabella 4.3 si riportano gli andamenti dei due modelli nella previsione della rete al tempo 33 rispetto alle metriche di classificazione. Da queste, emerge un andamento del modello non molto soddisfacente. Il tasso di errata classificazione è maggiore rispetto a quello rilevato per i precedenti istanti temporali, ma è comunque ancora accettabile. Emerge invece una percentuale elevata di falsi positivi. Una parte di questi archi classificati erroneamente è legata al fatto che ad ogni istante temporale si hanno degli attori che non presentano connessioni, per cui il modello stimerà degli archi tra nodi che invece risultano essere totalmente sconnessi, tuttavia questi eventi non sono prevedibili da parte del modello. In termini di confronto tra i due modelli, si può vedere che il secondo modello presenti un adattamento migliore del primo, in questo caso maggiormente amplificata rispetto al caso precedente.

	errata classificazione	*falsi positivi*	*falsi negativi*	*metrica F1*
$\mathbf{x}_t$ variabili, senza Z_t	0.373(0.025)	0.298(0.054)	0.668(0.051)	0.263(0.040)
$\mathbf{x}_t$ variabili, con Z_t	0.317(0.029)	0.216(0.050)	0.710(0.062)	0.269(0.032)

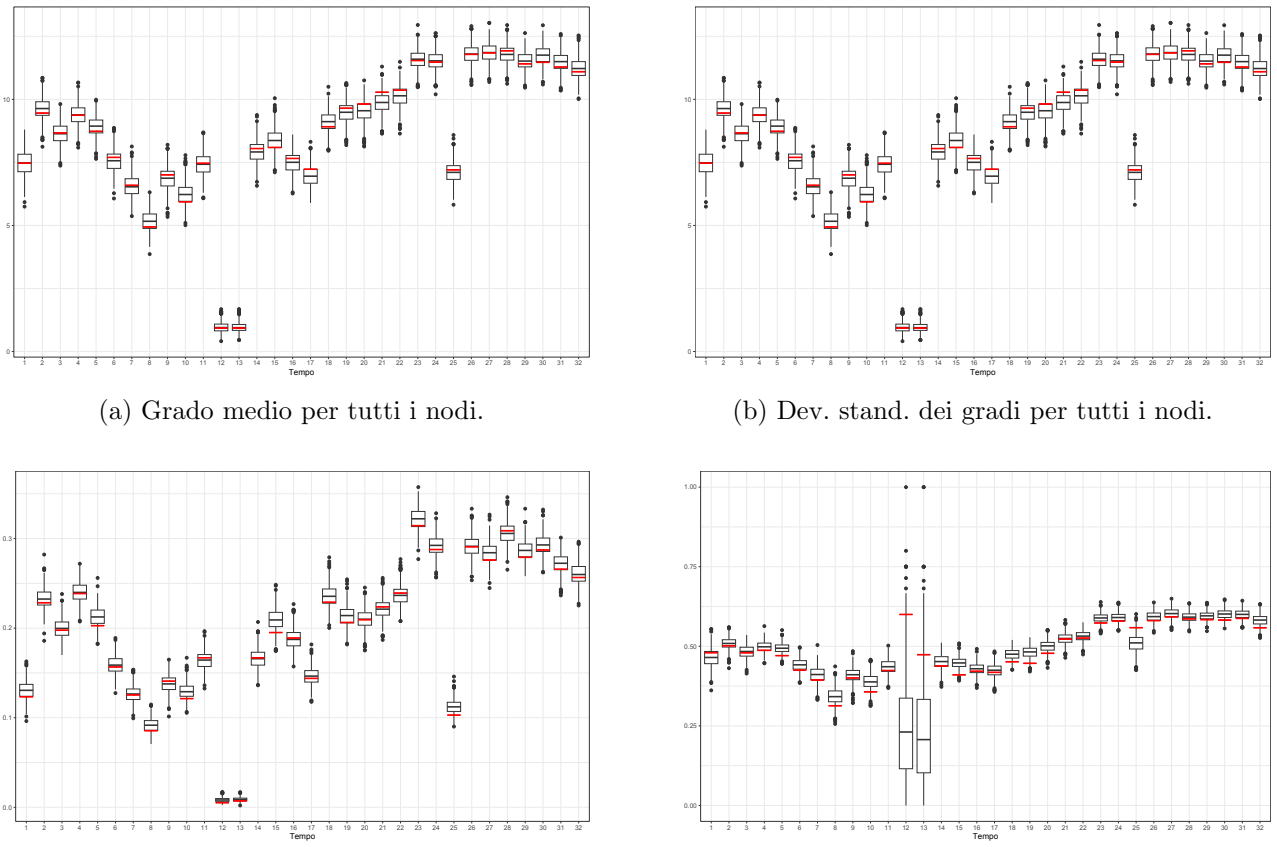
Tabella 4.3: Tassi di classificazione misurati nelle reti simulate dal modello a proiezioni latenti dinamico rispetto alla rete osservata al tempo 33.

	errata classificazione	*falsi positivi*	*falsi negativi*	*metrica F1*
$\mathbf{x}_t$ variabili, senza Z_t	0.101(0.001)	0.060(0.001)	0.273(0.005)	0.736(0.003)
$\mathbf{x}_t$ variabili, con Z_t	0.095(0.001)	0.060(0.002)	0.240(0.005)	0.760(0.003)
$\mathbf{x}_t$ costanti, senza Z_t	0.187(0.002)	0.123(0.003)	0.483(0.008)	0.530(0.004)
$\mathbf{x}_t$ costanti, con Z_t	0.182(0.002)	0.109(0.003)	0.453(0.008)	0.552(0.004)

Tabella 4.4: Tassi di classificazione per le quattro varianti considerate del modello a distanze latenti dinamico.

4.2 Applicazione: modello a distanze latenti dinamico

Si procede ora con l'analisi relative all'applicazione del modello a distanze latenti dinamico, presentato da Sewell & Chen (2015). Come nella sezione precedente, verrà di seguito proposto un confronto in termini di bontà d'adattamento tra i due modelli a posizioni latenti tempo-dipendenti con e senza variabili esplicative, verrà effettuato lo stesso confronto tra modelli con posizioni latenti costanti, e infine queste due sottoclassi verranno confrontate tra di loro, a seconda dell'impatto che si evidenzierà per le variabili esplicative. Le distribuzioni a priori considerate nei modelli con posizioni latenti variabili nel tempo per i parametri τ^2 e σ^2 hanno entrambe distribuzioni gamma inversa con iperparametri $a_\sigma = a_\tau = 2$, $b_\sigma = b_\tau = 1$, questa scelta consente imporre media pari a 1 e varianza non finita. Si usa dunque una distribuzione non informativa. In ogni caso, aumentando il valore dell'iperparametro di scala della distribuzione i risultati inferenziali non cambiano. Gli iperparametri scelti per γ a priori sono $\eta = 0.5$, $\xi = 1000$, poiché ci si aspetta un effetto positivo della propensione a formare un arco tra due nodi che presentano distanza euclidea inferiore a 1. La varianza è stata mantenuta elevata in modo da consentire a posteriori valori elevati. I coefficienti β associati alle variabili esplicative presentano tutti media nulla e varianza pari a 1000, in modo da consentire anche in questo caso valori elevati a posteriori, sia negativi sia positivi. Il numero di posizioni latenti selezionato con i criteri di informazione presentati in sezione 2.8 è $d = 9$ per il modello che non include variabili esplicative, mentre nell'altro caso è stato selezionato $d = 11$. Per i modelli con posizioni latenti costanti nel tempo, è stata scelta varianza a priori per ogni componente $\mathbf{x}_{i,h}$ pari a 5, mentre gli altri iperparametri non sono stati



(a) Grado medio per tutti i nodi.

(b) Dev. stand. dei gradi per tutti i nodi.

(c) Densità della rete.

(d) Transitività della rete.

Figura 4.2: Densità predittive a posteriori delle statistiche descrittive di nodo e di rete, calcolate rispetto ai primi 32 istanti temporali per il modello a distanze latenti dinamico. Le linee rosse presenti nei grafici corrispondono al valore della statistica calcolato rispetto alla rete osservata.

modificati. Per questa sottoclasse di modelli la dimensione della distribuzione associata alle posizioni latenti è stata scelta $d = 15$ per il modello senza variabili esplicative, $d = 16$ per il modello che invece le include.

I risultati dei modelli in termini di capacità previsiva sono presentati in Tabella 4.4. Per confrontare i quattro modelli si userà ancora la terminologia introdotta nella sezione precedente di modello 'completo' e 'ridotto'. Per entrambi i modelli completo e ridotto si evidenzia un miglioramento, seppur non molto elevato, portato dall'aggiunta delle variabili esplicative. Il tasso di falsi negativi è comunque abbastanza elevato per entrambi i sottomodelli. Si nota però che il modello ridotto presenta un tasso di errata classificazione e di falsi negativi quasi doppio rispetto al modello completo. Quest'ultimo sembra quindi più appropriato rispetto all'altro. A fronte anche del lieve miglioramento portato dalle variabili esplicative, il modello completo che include i termini Z_t sembra essere il modello migliore tra quelli presentati nella tabella. Verranno quindi mostrate le densità predittive a posteriori calcolate rispetto ad alcune proprietà delle reti (Figura 4.2) ottenute in corrispondenza del modello a coordinate

	Media	Deviazione Standard	Inferiore 0.05	Superiore 0.95
Chiamate effettuate	0.595	0.072	0.476	0.711
Messaggi scambiati	0.323	0.159	0.066	0.574
Anno accademico frequentato: GRT	-0.309	0.028	-0.355	-0.261
Anno accademico frequentato: Junior	0.113	0.112	0.061	0.165
Anno accademico frequentato: Senior	-0.501	0.029	-0.548	-0.454
Anno accademico frequentato: Secondo anno	0.130	0.014	0.106	0.153
Anno accademico frequentato: Sconosciuto	1.000	0.167	0.733	1.280
Piano del dormitorio: 289.2	-0.212	0.033	-0.267	-0.156
Piano del dormitorio: 289.3	-0.050	0.018	-0.082	-0.019
Piano del dormitorio: 289.4	-0.855	0.028	-0.903	-0.806
Piano del dormitorio: 290.1	-0.059	0.110	-0.232	0.122
Piano del dormitorio: 290.2	0.055	0.054	-0.029	0.132
Piano del dormitorio: 290.3	0.019	0.020	-0.023	0.059
Piano del dormitorio: 290.4	-1.080	0.057	-1.170	-0.984
Piano del dormitorio: Sconosciuto	-1.360	0.064	-1.460	-1.260
Residenza nello stesso piano	1.110	0.025	1.070	1.159
Stesso anno accademico frequentato	-0.227	0.025	-0.270	-0.188

Tabella 4.5: Coefficienti delle variabili esplicative associate al modello a distanze latenti dinamico con posizioni latenti tempo-dipendenti.

latenti tempo-dipendenti con variabili esplicative.

Le distribuzioni predittive a posteriori mostrate in Figura 4.2 evidenziano un buon adattamento del modello rispetto alla rete osservata. Infatti, tali densità predittive risultano essere coerenti per tutte e quattro le statistiche indagate, rispetto al valore realmente osservato nella rete dinamica. Anche rispetto alle simulazioni, sembra che il livello di transitività sia abbastanza coerente rispetto a quello osservato, sopratutto a partire dal tempo 14, tempo dal quale sembra esserci meno variabilità nelle distribuzioni ottenute, rispetto ai tempi precedenti. Per tutte le altre distribuzioni si evidenzia un elevata variabilità per tutti i tempi considerati, fatta eccezione per i tempi 12 e 13 dove però la rete ha una densità molto bassa. Rispetto alle stesse densità predittive del modello a proiezioni latenti (Figura 4.1) le statistiche sembrano essere più coerenti, ovvero sembrano essere maggiormente orientate nei valori osservati nella rete dinamica reale. Inoltre, in termini di variabilità delle distribuzioni i due modelli sembrano essere equivalenti.

Avendo selezionato il modello con coordinate latenti tempo-dipendenti con variabili esplicative,

vengono di seguito presentati i coefficienti a posteriori ottenuti in fase di modellazione (Tabella 4.5). Tali coefficienti esprimono l'effetto sul logaritmo del rapporto delle quote (*log odds*) rispetto ad un incremento unitario di una variabile esplicativa, fermo restando le altre. L'interpretazione è quindi la stessa fornita per i coefficienti del modello mostrati nella sezione precedente. Le stime puntuali a posteriori e gli intervalli di credibilità non si discostano in modo evidente da quelle mostrate in Tabella 4.2 per quanto riguarda l'effetto delle chiamate e messaggi effettuati e per l'effetto rispetto allo stesso anno accademico e stesso piano di residenza, mentre sembra esserci un effetto diverso per le variabili esplicative relative all'anno di corso frequentato e al piano di residenza. Si notano infatti alcuni coefficienti non significativi per il piano di residenza (tutti quelli relativi a 290), e alcuni coefficienti relativi all'anno frequentato presentano un effetto stimato con un segno diverso, quindi il tipo di effetto (positivo o negativo) differisce tra i due modelli. Un esempio può essere l'anno accademico degli Junior e dei ragazzi del secondo anno, che nel precedente modello assumevano un effetto negativo rispetto alla modalità di riferimento.

Si passa ora alla valutazione dell'andamento dei modelli con posizioni latenti tempo-dipendenti con e senza variabili esplicative, a fronte di quanto emerso prima, per la previsione di una rete futura. Questi due modelli vengono ancora confrontati per verificare se l'inclusione delle variabili esplicative possa ancora portare un miglioramento in termini di bontà d'adattamento, e in caso di risposta affermativa, si vuole verificare di quanto vari. Le assunzioni distributive e gli iperparametri scelti non vengono modificati rispetto a quanto specificato prima.

In Tabella 4.6 si riportano i tassi di classificazione ottenuti. Il tasso di errata classificazione è abbastanza soddisfacente. Si nota però un elevato tasso di falsi negativi, il modello sottostima in modo forte il numero di connessioni della rete, mentre commette un errore minimo in termini di falsi positivi, per cui gli archi previsti sono effettivamente corretti. Questo andamento comporta un valore molto basso della metrica F1. Il confronto dei due modelli evidenzia che l'inclusione delle variabili esplicative sembra portare una riduzione del tasso di falsi negativi, il cui valore rimane comunque alto, ma in termini di metrica F1 si rileva un lieve miglioramento.

Un possibile approccio alternativo per mantenere lo stesso andamento temporale rilevato fino al

	errata classificazione	*falsi positivi*	*falsi negativi*	*metrica F1*
$\mathbf{x}_t$ variabili, senza Z_t	0.187(0.006)	0.014(0.006)	0.867(0.039)	0.221(0.055)
$\mathbf{x}_t$ variabili, con Z_t	0.183(0.006)	0.015(0.006)	0.836(0.040)	0.267(0.053)

Tabella 4.6: Tabella dei tassi di classificazione misurati nelle reti simulate dal modello a distanze latenti dinamico rispetto alla rete al tempo 33.

	errata classificazione	*falsi positivi*	*falsi negativi*	*metrica F1*
$\mathbf{x}_t$ variabili, senza Z_t	0.098(0.001)	0.060(0.001)	0.256(0.002)	0.745(0.002)
$\mathbf{x}_t$ variabili, con Z_t	0.144(0.001)	0.087(0.002)	0.383(0.002)	0.624(0.002)
$\mathbf{x}_t$ costanti, senza Z_t	0.197(0.002)	0.122(0.003)	0.511(0.006)	0.490(0.004)
$\mathbf{x}_t$ costanti, con Z_t	0.194(0.002)	0.117(0.003)	0.517(0.006)	0.490(0.004)

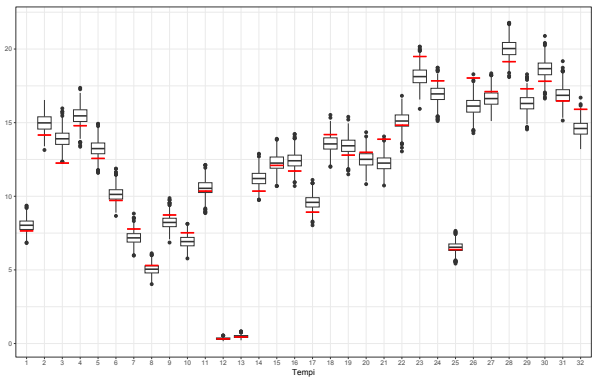
Tabella 4.7: Tassi di classificazione di quattro varianti del modello a distanze latenti con mistura.

tempo 32 può essere quello di classificare come 0 e 1 le probabilità di connessione previste ottenute se queste risultano minori o maggiori di una certa soglia, selezionata in modo che assicuri una densità della rete pari a quelle rilevate, per esempio, come media della densità di tutte le reti disponibili. Questo può essere fatto imponendo come valore della soglia il quantile $1 - \bar{D}$ della distribuzione delle probabilità previste per ogni cella della matrice di adiacenza, ovvero probabilità ottenute rispetto a tutti i campioni disponibili. Questo approccio potrebbe portare a delle previsioni migliori, ma per consistenza rispetto alle analisi proposte anche per gli altri modelli, questo metodo non è stato approfondito.

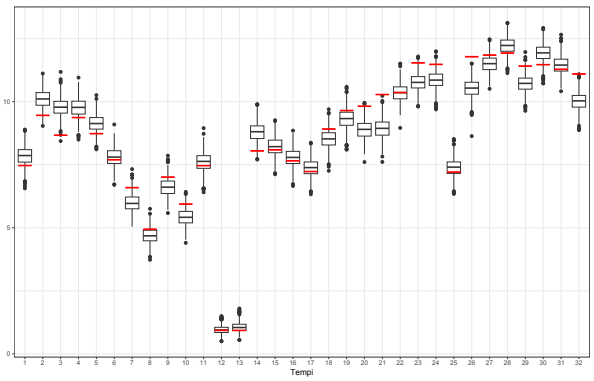
4.3 Applicazione: modello a distanze latenti con mistura

Si passa infine alla descrizione dell'analisi effettuata per l'ultimo modello proposto. Le distribuzioni a priori scelte per i vettori d-dimensionali delle G medie consistono in distribuzioni normali a media nulla, e con varianza pari a 1000 per ogni componente, mantenendo indipendenza tra le stesse. La varianza σ_g^2 per ogni gruppo ha distribuzione a priori chi quadro inversa con gradi di libertà $\alpha = 2$, in modo che la distribuzione fosse di tipo non informativo. Infatti, per tale scelta il parametro non ha media e varianza finita a priori. Il parametro σ_0^2 è stato scelto pari a 1, in modo da non apportare variazioni di scala. I parametri λ hanno distribuzione a priori di Dirichlet con iperparametri $\nu_1 = \frac{1}{G}, \ldots, \nu_G = \frac{1}{G}$. Il parametro β associato alle variabili esplicative ha distribuzione a priori normale multivariata con media nulla e varianza pari a 1000 per ogni componente, mantenendo indipendenza tra di esse. Infine, per il parametro γ associato alle posizioni latenti è stata scelta media a priori 0.5 poiché ci si aspetta che esso sia positivo, e varianza pari a 1000. Tramite i criteri di informazione *loo*, WAIC e *pseudo-BMA* sono stati selezionati in modo congiunto il numero di dimensioni delle posizioni latenti e il numero di gruppi, scegliendo $d = 6$ e $G = 3$ per tutte le quattro varianti del modello considerate.

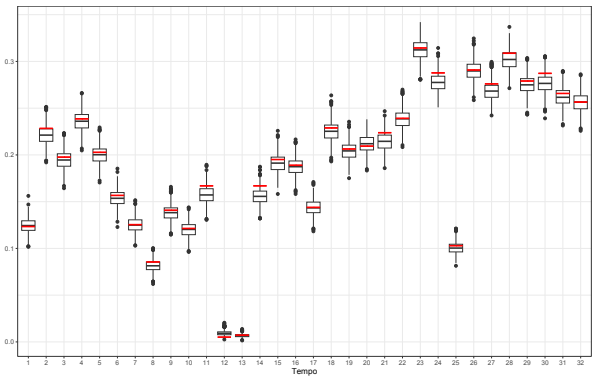
Come per i casi precedenti, verrà mostrata prima la bontà d'adattamento del modello rispetto alle 32 reti rilevate, si valuterà la bontà della previsione la rete futura, e infine si analizzeranno i *cluster* identificati dal modello. I tassi di classificazione risultanti per i quattro modelli considerati (Tabella 4.7) portano a preferire il modello senza variabili esplicative con $\mathbf{x}_t$ variabili, in contrasto con quanto



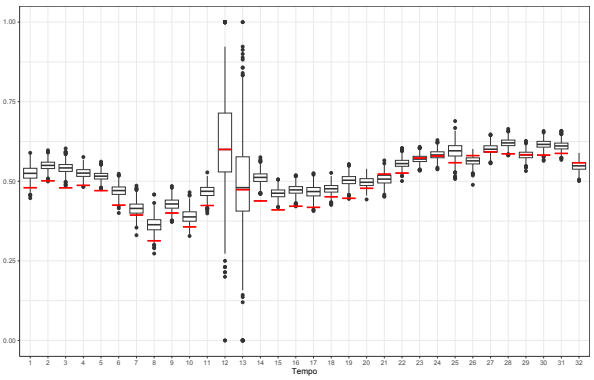
(a) Grado medio per tutti i nodi.



(b) Dev. stand. dei gradi per tutti i nodi.



(c) Densità della rete.



(d) Transitività della rete.

Figura 4.3: Densità predittive a posteriori delle statistiche descrittive di nodo e di rete, calcolate rispetto ai primi 32 istanti temporali per il modello a distanze latenti con mistura. Le linee rosse presenti nei grafici corrispondono al valore della statistica calcolato rispetto alla rete osservata.

fatto per i modelli precedenti. Infatti, si evince un tasso di falsi negativi maggiore per il modello con i termini Z_t rispetto all'altro, le variabili esplicative portano quindi a sottostimare il numero di archi presenti entro la rete. Un andamento di questo tipo può essere legato al fatto che le variabili esplicative considerate sono tutte di natura qualitativa, e possono anch'esse essere assimilabili a gruppi d'appartenenza dei soggetti. Introdurre quindi altra informazione legata a gruppi porta ad un peggioramento nell'identificazione dei *cluster* entro al modello. I modelli a coordinate latenti tempo-dipendenti presentano comunque un adattamento migliore rispetto a quelli con $\mathbf{x}_t$ costanti, in accordo con i risultati precedenti.

In Figura 4.3 sono mostrati gli andamenti delle distribuzioni predittive a posteriori delle statistiche descrittive di rete e di nodo. Osservando la Figura 4.3a si può notare che il grado medio in corrispondenza dei primi istanti temporali viene sovrastimato, per i tempi centrali le distribuzioni predittive sono coerenti con quanto osservato, ma per gli ultimi istanti temporali il modello non riesce a identificare correttamente tale statistica, sottostimando la media dei gradi. Questo fenomeno si può osservare

	errata classificazione	*falsi positivi*	*falsi negativi*	*metrica F1*
$\mathbf{x}_t$ variabili, senza Z_t	0.393(0.027)	0.349(0.044)	0.563(0.048)	0.312(0.040)
$\mathbf{x}_t$ variabili, con Z_t	0.310(0.023)	0.205(0.041)	0.721(0.066)	0.267(0.040)

Tabella 4.8: Tassi di classificazione dei modelli a distanze latenti con mistura per la previsione della rete al tempo 33.

anche per la densità media (Figura 4.3c) ma in modo meno evidente. Il livello di transitività (Figura 4.3d) viene sovrastimato per quasi tutti gli istanti temporali, le distribuzioni predittive a posteriori non sono centrate rispetto alla transitività misurata nella rete osservata. Come nei casi precedenti, le densità predittive hanno distribuzioni molto variabili. La deviazione standard dei gradi (Figura4.3b) non viene correttamente identificata dal modello, presentando distribuzioni non coerenti con quelle osservate.

Si può ora passare all'analisi della previsione di una nuova rete. A tale scopo, sono state calcolate le probabilità a posteriori d'appartenenza alle G classi selezionate per ogni nodo, in modo da simulare la classe d'appartenenza di ogni nodo da una distribuzione multinomiale con delle probabilità specifiche per ogni nodo. Dopo aver simulato la classe d'appartenenza $\widehat{K_i} = g$, il valore della coordinata latente viene simulato dalla distribuzione $N_d(\widetilde{\upsilon_g}, \widetilde{\sigma_g^2} I)$, dove con $\widetilde{\upsilon_g}$ e $\widetilde{\sigma_g^2}$ ci si riferisce alla media e varianza a posteriori. A questo punto, si può calcolare il predittore lineare e simulare da una distribuzione di Bernoulli gli archi della rete futura. I risultati ottenuti per i due modelli a posizioni latenti variabili nel tempo (Tabella 4.8) non risultano affatto soddisfacenti. Il modello senza variabili esplicative presenta un tasso di falsi negativi inferiore rispetto al modello con variabili esplicative e una metrica F1 maggiore, tuttavia presenta tasso di errata classificazione e tasso di falsi positivi maggiore. Perciò, anche da questa analisi il modello che risulta preferibile è quello che non include le variabili esplicative.

A fronte dei risultati ottenuti sia rispetto ai dati usati per adattare il modello sia in fase di previsione, sembra ragionevole svolgere l'analisi dell'appartenenza ai gruppi sulla base delle assegnazioni svolte dal modello che include le variabili esplicative. Come detto prima, il numero di gruppi identificato dal modello è $G = 3$, con proporzioni d'appartenenza $0.32, 0.17, 0.51$. I gruppi identificati dal modello non evidenziano una particolare propensione rispetto alle variabili esplicative dell'anno accademico frequentato o del piano di residenza, ma sembrano invece differire in termini di numero di connessioni entro il gruppo, e sembra che il grado medio di ogni nodo rispetto a tutti i 32 istanti temporali, in riferimento all'intera rete e non al gruppo d'appartenenza, assuma dei valori diversi a seconda del gruppo. In Figura 4.4 vengono mostrate le densità predittive a posteriori della densità di connessioni delle comunità riconosciute dal modello. Si può vedere come i nodi del primo grup-

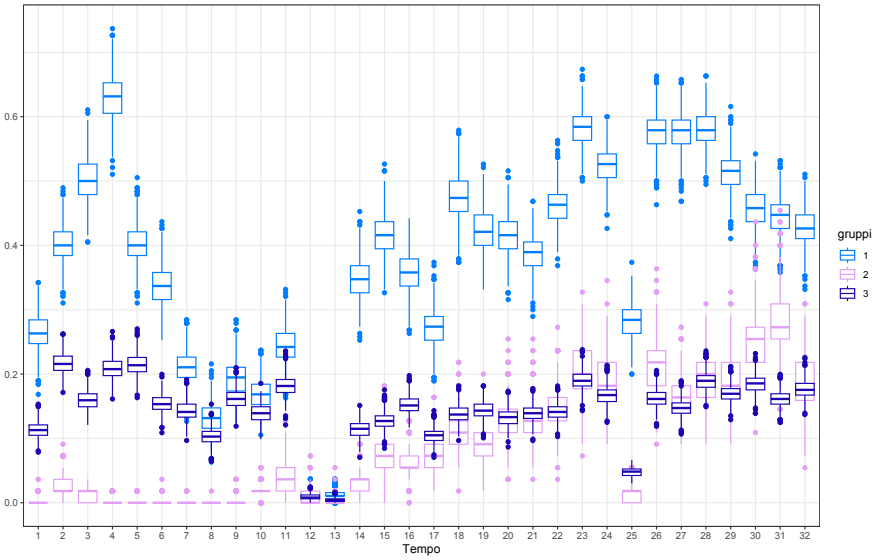
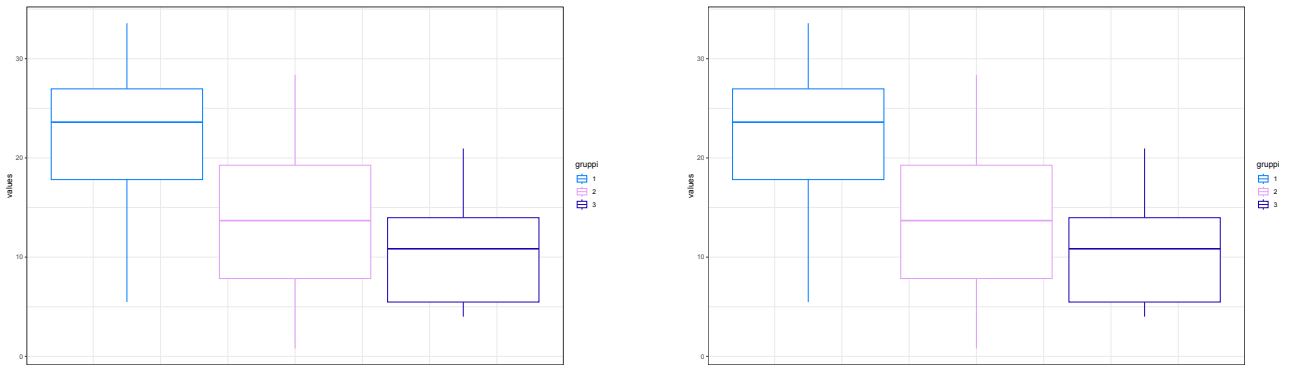


Figura 4.4: Densità medie per i tre gruppi identificati dal modello nei 32 tempi considerati.

po presentino un numero di connessioni entro la comunità molto maggiore rispetto agli altri due, il secondo gruppo sembra assumere un comportamento medio, che rimane stabile per tutta la durata del periodo d'osservazione, mentre il terzo gruppo inizialmente non ha connessioni entro il gruppo, mentre dal tempo 14 in poi questo valore aumenta, tanto che il secondo e terzo gruppo sembrano assumere un andamento molto simile. Sembra quindi che dopo il tempo 14 il numero di gruppi entro la rete dinamica passi da 3 a 2. Questo tipo di analisi non è stato approfondito, poiché richiederebbe lo sviluppo di un modello che consentisse flessibilità temporale in termini di appartenenza al gruppo e numero di gruppi.

Un'analisi ulteriore dell'andamento dei tre gruppi porta a concludere che i vertici appartenenti al primo gruppo presentano un numero di connessioni maggiore rispetto agli altri vertici, non solo entro la comunità stessa ma in termini globali. Inoltre, considerando in modo separato il numero di connessioni prima e dopo il periodo di Natale (che corrisponde ai tempi 12 e 13) emerge che il terzo gruppo presenta un numero di connessioni molto ridotto nel primo periodo, e la differenza tra i tre gruppi è molto accentuata (Figura 4.5). Dal tempo 14 in poi, i nodi del terzo gruppo presentano un numero di connessioni maggiore, ma comunque inferiore rispetto agli altri due. Sembra quindi plausibile che il modello riesca a distinguere un gruppo di studenti che presenta un elevato numero di connessioni sopratutto tra i soggetti dello stesso gruppo, un gruppo di soggetti che inizialmente non presentava connessioni e un gruppo con un comportamento medio.



(a) Grado medio registrato tra gli istanti 1 e 13.

(b) Grado medio registrato tra gli istanti 14 e 32.

Figura 4.5: Distribuzione del grado medio per i nodi suddivisi nelle comunità identificate dal modello a distanze latenti con mistura, calcolati nella rete realmente osservata.

4.4 Commenti finali sull'analisi

Dopo aver valutato nel dettaglio le bontà d'adattamento dei modelli, è possibile operare un confronto tra di essi. Il modello che presenta prestazioni migliori in termini di tassi di classificazione ottenuti rispetto ai dati usati per adattare i modelli, di densità predittive a posteriori delle statistiche di rete e di vertici considerate e anche di previsione è il modello a distanze latenti dinamico. Il modello a proiezioni latenti dinamico presenta comunque un andamento simile in termini di tassi di classificazione rispetto alle 32 reti usate per adattare il modello e densità a posteriori delle statistiche considerate, ma commette un tasso di errata classificazione maggiore in termini previsivi. Questo modello risulta infatti troppo poco conservativo, classificando molti archi come presenti nella rete, che poi non si realizzano. In questo senso, si preferisce un modello che classifica meno archi come successi rispetto a quelli realmente presenti, ma non commette errori tra quelli che classifica come presenti, portando quindi ad avere pochi archi ma ad assegnazioni tutte corrette.

Il modello a distanze latenti con mistura porta a dei risultati soddisfacenti in termini di tassi di classificazione rispetto alle reti con cui è stato adattato il modello e trova dei gruppi sensati e coerenti, ma in termini previsivi non ha una buona prestazione. In ogni caso, è in grado di evidenziare aspetti interessanti rispetto all'evoluzione della rete dinamica, nonostante si consideri un solo gruppo d'appartenenza fissato per tutti i tempi. In termini di socialità degli studenti, si può evincere un'evoluzione nel comportamento dei soggetti, da quelli che in un anno accademico sono stati in grado di accrescere il numero di persone con cui si sono incontrati, ai soggetti che non hanno modificato le loro abitudini

in termini di socialità 'fisica'. Il gruppo che presenta relazioni nulle con i soggetti può essere spiegato in tre modi distinti: un'interpretazione può essere legata al fatto che quei soggetti per qualche ragione non hanno partecipato inizialmente allo studio e si sono inseriti a partire da Gennaio, oppure per qualche ragione non erano presenti nel campus, oppure ancora effettivamente sono rimasti isolati per un elevato periodo di tempo.

Conclusioni

In questo elaborato di tesi sono stati utilizzati dei modelli a spazi latenti per analizzare una rete sociale dinamica che misura la prossimità tra gli studenti di un dormitorio del MIT entro una cadenza settimanale. L'obiettivo dell'analisi era di comprendere i meccanismi sottostanti alla formazione degli archi e cogliere le dinamiche dell'evoluzione temporale della rete. A tale scopo, sono state utilizzate 32 delle 37 reti a disposizione per adattare i modelli considerati, ed è stata effettuata la previsione della rete al tempo 33. I modelli a spazi latenti utilizzati, quali il modello a proiezioni latenti dinamico di Durante & Dunson (2014) e il modello a distanze latenti dinamico di Sewell & Chen (2015) si sono rivelati entrambi in grado di cogliere la co-evoluzione temporale delle reti, individuare in modo soddisfacente le proprietà di nodo e di rete, oltre a classificare correttamente gli archi delle reti. Attraverso il modello a distanze latenti con mistura di Handcock et al. (2007) è stato possibile identificare delle comunità entro la rete, mantenendo l'assunzione di appartenenza al medesimo gruppo, da parte dei nodi, per tutti gli istanti temporali. Il modello ha identificato tre gruppi entro la rete, operando una divisione tra nodi più e meno attivi. Questo modello non ha però raggiunto le prestazioni degli altri due in termini di classificazione degli archi e capacità di identificare le proprietà della rete. Una possibile estensione rispetto alle analisi eseguite è quella di considerare un modello che rilevi la presenza di gruppi entro una rete dinamica più flessibile di quello utilizzato, considerando per esempio un numero di gruppi e l'appartenenza dei nodi al gruppo variabile nel tempo, che potrebbe portare dei buoni risultati, come emerso dalle presenti analisi. Altre possibili estensioni riguardano l'analisi della rete dinamica utilizzata non come rete binaria ma bensì come rete pesata, considerando il numero di volte in cui gli studenti si sono incontrati nell'arco di una settimana. Inoltre, sempre come possibile ulteriore analisi, si potrebbe valutare in termini previsivi l'inclusione di un nuovo nodo entro la rete e prevedere le sue connessioni, sia rispetto ad una rete in corrispondenza di un determinato tempo, sia rispetto alla rete dinamica nel suo complesso, valutando quindi la sua co-evoluzione con gli altri nodi.

Bibliografia

Bavelas, A. (1948). A mathematical model for group structures. *Human organization* **7**, 16–30.

Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association* **112**, 859–877.

Durante, D. & Dunson, D. B. (2014). Nonparametric bayes dynamic modelling of relational data. *Biometrika* **101**, 883–898.

Frank, O. & Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association* **81**, 832–842.

Freeman, L. (2004). The development of social network analysis. *A Study in the Sociology of Science* **1**, 159–167.

Fruchterman, T. M. & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and experience* **21**, 1129–1164.

Handcock, M. S., Raftery, A. E. & Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **170**, 301–354.

Heider, F. (1946). Attitudes and cognitive organization. *The Journal of psychology* **21**, 107–112.

Hoff, P. D., Raftery, A. E. & Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the american Statistical association* **97**, 1090–1098.

Holland, P. W. & Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association* **76**, 33–50.

Hunter, D. R., Krivitsky, P. N. & Schweinberger, M. (2012). Computational statistical methods for social network models. *Journal of Computational and Graphical Statistics* **21**, 856–882.

Kamada, T., Kawai, S. et al. (1989). An algorithm for drawing general undirected graphs. *Information processing letters* **31**, 7–15.

Karinthy, F. (1929). Chain-links. *Everything is different* , 21–26.

Kilduff, M., Liu, L. & Tasselli, S. (2023). A connected world: Social networks and organizations. *Elements in Organization Theory* .

Kolaczyk, E. D. & Csárdi, G. (2014). *Statistical analysis of network data with R*, vol. 65. Springer.

Krivitsky, P. N. & Handcock, M. S. (2008). Fitting position latent cluster models for social networks with latentnet. *Journal of statistical software* **24**.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A. & Blei, D. M. (2016). Automatic differentiation variational inference.

Lewin, K. (1936). A dynamic theory of personality: Selected papers. *The Journal of Nervous and Mental Disease* **84**, 612–613.

Lorrain, F. & White, H. C. (1971). Structural equivalence of individuals in social networks. *The Journal of mathematical sociology* **1**, 49–80.

Madan, A., Cebrian, M., Moturu, S., Farrahi, K. et al. (2011). Sensing the" health state" of a community. *IEEE Pervasive Computing* **11**, 36–45.

Milgram, S. (1967). The small world problem. *Psychology today* **2**, 60–67.

Moreno, J. L. (1934). *Who shall survive?: A new approach to the problem of human interrelations.* Nervous and mental disease publishing co.

Nowicki, K. & Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association* **96**, 1077–1087.

Sarkar, P. & Moore, A. W. (2005). Dynamic social network analysis using latent space models. *Acm sigkdd explorations newsletter* **7**, 31–40.

Sewell, D. K. & Chen, Y. (2015). Latent space models for dynamic networks. *Journal of the American Statistical Association* **110**, 1646–1657.

van Duijn, M. (1995). *"Estimation of a Random Effects Model for Directed Graphs."*. pp. 113 – 132.

van Duijn, M., Snijders, T. & Zijlstra, B. (2004). P2: A random effects model with covariates for directed graphs. *Statistica Neerlandica* **58**, 234–254.

Vehtari, A., Gelman, A. & Gabry, J. (2016). Practical bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* **27**, 1413–1432.

Wasserman, S. & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge university press.

Yao, Y., Vehtari, A., Simpson, D. & Gelman, A. (2018). Using stacking to average bayesian predictive distributions (with discussion). *Bayesian Analysis* **13**.

Zijlstra, B. J. H., van Duijn, M. A. J. & Snijders, T. A. B. (2009). Mcmc estimation for the p2 network regression model with crossed random effects. *British Journal of Mathematical and Statistical Psychology* **62**, 143–166.