

UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA
SCUOLA DI ECONOMIA E STATISTICA

CORSO DI LAUREA IN
SCIENZE STATISTICHE ED ECONOMICHE



ANALISI DELLA STRUTTURA INTERNA
DE LA LOMBARDIA, ATTRAVERSO
L'UTILIZZO DI MODELLI A BLOCCHI
STOCASTICI ESTESI

RELATORE: Prof. Tommaso Rigon

CORRELATORE: Prof. Bernardo Nipoti

TESI DI LAUREA DI:
Luna Cantaroni
MATRICOLA N. 905193

ANNO ACCADEMICO 2024/2025

Indice

Introduzione	1
1 Sulla rete criminale <i>La Lombardia</i>	3
1.1 <i>Operazione Infinito</i>	3
1.2 Analisi descrittiva dei dati	5
2 Modelli probabilistici	13
2.1 Concetti preliminari	13
2.1.1 Metodi Markov Chain Monte Carlo	15
2.1.2 Metodi bayesiani non parametrici	17
2.1.3 Modelli di mistura	18
2.1.4 Distribuzioni a priori non parametriche discrete: classe di tipo Gibbs	21
2.1.5 Stirling-Gamma	31
2.1.6 SALSO	34
2.2 Modelli per reti	39
2.2.1 Modelli per blocchi stocastici	39
2.2.2 Modelli per blocchi stocastici estesi	43
3 Valutazione dei ESBM attraverso dati simulati	49
3.1 Applicazione su reti simulate semplici	49
3.2 Applicazione su rete che simula le caratteristiche della <i>rete Infinito</i> . .	80
4 Applicazione dei ESBM alla <i>rete Infinito</i>	85
4.1 Performance dei ESBM applicati alla <i>rete Infinito</i> senza l'inclusione degli attributi dei nodi	85
4.2 Performance dei ESBM applicati alla <i>rete Infinito</i> includendo gli attributi dei nodi	86

4.3 Analisi della struttura interna de <i>La Lombardia</i> attraverso il ESBM con distribuzione a priori di Gneden supervisionata	87
Conclusione	93
Bibliografia	95

Introduzione

La 'Ndrangheta è un'organizzazione criminale di stampo mafioso, nata in Calabria, basata sulla famiglia di sangue e organizzata in 'ndrine. La 'ndrina (dal dialetto reggino “ciò che non si piega”) costituisce l'elemento organizzativo fondamentale, nella quale si può entrare a farne parte solo attraverso un rituale d'affiliazione. Le 'ndrine appartenenti ad una stessa area formano una *locale*, che assume il controllo di uno specifico territorio. La *locale* è la principale unità strutturale della 'Ndrangheta, al suo interno ci sono diverse cariche formali, ognuna con una specifica funzione: il boss detto “capolocale” è responsabile dei fondi comuni della *locale*, il “crimine” controlla le azioni violente, mentre il “mastro di giornata” si occupa dei flussi di comunicazione interni.

A partire dagli anni '90 del secolo scorso la 'Ndrangheta ha saputo espandersi sia a livello nazionale che internazionale, sul nostro territorio è risultata particolarmente rilevante la sua crescita nell'area lombarda, tanto da assumere il nome *La Lombardia*. Un'importante indagine, denominata *Operazione Infinito*, è stata condotta dalle forze dell'ordine italiane tra il 2007 e il 2009 proprio in tale regione, per portarne alla luce l'operato. Grazie ai dati raccolti durante questa operazione, è possibile analizzare la struttura interna de *La Lombardia*.

Questa organizzazione criminale è caratterizzata da “blocchi” di individui che interagiscono tra di loro secondo una complessa combinazione di gerarchie, verticali e orizzontali. Questa struttura, caratterizzata anche da nuclei periferici e architetture difformi, risulta particolarmente ostica da dipanare, per questo motivo gli algoritmi di rilevamento di comunità classici, che fanno inferenza su gruppi di criminali caratterizzati da una densa connettività all'interno dei blocchi e da connessioni meno forti tra gruppi differenti, non riescono a rilevare tutte le strutture dei blocchi, trascurandone anche di importanti, come nuclei periferici e schemi di comunità difformi e deboli. Anche gli algoritmi di clustering spettrali e i modelli per blocchi stocastici, nonostante riescano ad imparare architetture tra blocchi più generali, non riescono a cogliere tutti gli aspetti che caratterizzano le strutture criminali.

Per riuscire dunque a rilevare la coesistenza di svariate comunità, nuclei periferici e architetture difformi, di cui non è noto il numero, la dimensione e la struttura, anche a causa degli errori di misurazione che scaturiscono dalle indagini, [Legramanti et al. \(2022\)](#) hanno sviluppato un'estensione dei modelli per blocchi stocastici (ESBM).

In questo elaborato perciò, andremo ad utilizzare tali modelli per scoprire la struttura interna de *La Lombardia*, introducendo alcuni elementi di innovazione rispetto agli studi portati avanti in precedenza. In primo luogo andremo a caratterizzare i dati a disposizione e condurremo un'analisi descrittiva per dati di rete volta ad evidenziarne le caratteristiche principali. In secondo luogo andremo ad illustrare i modelli che utilizzeremo per la costruzione di un modello per blocchi stocastici estesi, in particolare descriveremo prima quelli relativi all'approccio bayesiano non parametrico e successivamente quelli che trattano dati di rete. In seguito mostreremo il corretto funzionamento del modello sviluppato attraverso l'utilizzo di dati simulati e infine lo utilizzeremo sui dati ottenuti dall'*Operazione Infinito* per studiare la struttura interna de *La Lombardia*.

Capitolo 1

Sulla rete criminale *La Lombardia*

In questo capitolo ci occuperemo dei dati raccolti durante l'*Operazione Infinito*, i quali ci permetteranno di caratterizzare la struttura interna della *La Lombardia*. Andremo prima di tutto a vedere come questi sono stati raccolti e ripercorreremo i passaggi effettuati per trasformarli in dati di rete. In seguito, procederemo con un'analisi descrittiva, per studiare le caratteristiche principali di questa rete criminale.

1.1 *Operazione Infinito*

L'*Operazione Infinito* è un'operazione portata avanti dalle forze dell'ordine italiane tra il 2007 e il 2009, volta a penetrare all'interno della base lombarda della 'Ndrangheta, così da dipanare e sabotare la sua architettura, con un'attenzione particolare nel tracciarne la struttura gerarchica e nell'individuazione delle differenti *locali* esistenti nella regione. Le indagini hanno riguardato più di duecento persone, ed i reati contestati vanno dall'omicidio, al traffico di sostanze stupefacenti, dal riciclaggio di denaro proveniente dalle attività illecite quali corruzione, estorsione ed usura all'ostacolo al libero esercizio del diritto di voto. Grazie a queste indagini gli inquirenti hanno potuto tracciare un censimento delle *locali* della 'Ndrangheta in Lombardia, individuando i 15 nuclei territoriali maggiormente esposti: Milano, Cormano, Bollate, Bresso, Corsico, Legnano, Limbiate, Solaro, Pioltello, Rho, Pavia, Canzo, Mariano Comense, Erba, Desio e Seregno.

I documenti relativi all'ordine di detenzione pre processuale emesso dal giudice delle investigazioni preliminari su richiesta della procura, forniscono informazioni su innumerevoli incontri tra affiliati, sia avvenuti in privato che in luoghi pubblici. I dati su cui lavoreremo provengono dal sito <https://sites.google.com/site/ucinetsoftware/datasets/covert-networks>, questi descrivono la partecipazione di 156 sospettati a 47 diversi incontri dell'organizzazione criminale monitorati,

secondo quanto riportato negli atti giudiziari del Tribunale di Milano. Di questi solo 118 hanno preso parte ad almeno un summit e sono stati classificati dagli atti come appartenenti ad una specifica organizzazione criminale, ci limiteremo quindi a considerare solo questi affiliati. La matrice di partenza ha dunque dimensione 156×47 , è tale che ogni riga corrisponde ad un individuo e ogni colonna a un summit differente, ogni cella è caratterizzata dal numero 1 o 0, dove il primo indica la presenza dell'individuo all'incontro e il secondo la sua assenza. Analizzando la co-partecipazione degli individui ai summit monitorati, è stato rilevato che solo il 18% delle coppie di affiliati ha assistito ad almeno un incontro in comune e solo il 5% a più di uno. Questo riflette la complessità dell'organizzazione criminale e le misure di coordinamento altamente regolamentate adottate dalla 'Ndrangheta per mantenere la propria sicurezza, per questo motivo possiamo ritenere sufficiente la partecipazione da parte di due sospettati diversi ad uno stesso incontro per stabilire una connessione tra i due.

Dalla matrice iniziale ne abbiamo poi creato una detta matrice di adiacenza, questa è binaria e simmetrica, inoltre ogni riga e ogni colonna corrisponde ad un affiliato e le celle rappresentano la partecipazione dei due individui considerati ad almeno un incontro in comune (1) o il non avvenimento di tale fatto (0). Come conseguenza della forma dispersa e quasi binaria dei conteggi originali, questa dicotomizzazione porta ad una perdita trascurabile dell'informazione, inoltre è utile per ridurre il rumore che potrebbe scaturire dalle investigazioni. Più problematica è la possibile presenza di falsi negativi, che potrebbe sorgere dalle strategie di copertura messe in atto dall'organizzazione criminale. Questa procedura di dicotomizzazione non altera la struttura interna dell'organizzazione e motiva oltretutto l'inclusione di conoscenza esterna, rispetto a quella contenuta nella matrice di adiacenza, su *La Lombardia* e sugli affiliati stessi.

Le attuali teorie forensi e le prime analisi quantitative effettuate su tale organizzazione criminale, suggeriscono che la composizione interna della 'Ndrangheta ruoti attorno a specifiche relazioni di sangue, che potrebbero ulteriormente essere aggregate a livello territoriale in unità dette *locali*. Ciascuna *locale* controlla uno specifico territorio ed è regolata da gerarchie interne, come la 'Ndrangheta stessa; è infatti composta da un gruppo di affiliati e da pochi boss che hanno il compito di guidarla. Gli atti giudiziari forniscono informazioni sul ruolo all'interno dell'organizzazione, boss o affiliato, e la *locale* d'appartenenza. Queste informazioni esterne, se integrate nel modello utilizzato per fare inferenza, potrebbero aiutare a rilevare la struttura a blocchi sottostante all'organizzazione criminale, riducendo l'impatto di strategie

nascoste. È importante però, tenere in considerazione che questo tipo d'informazione può essere soggetta ad un errore di misurazione. È possibile la valutazione di tale errore di misurazione attraverso l'utilizzo di modelli per blocchi stocastici estesi (Legramanti et al., 2022) che ricorrono a prior di tipo Gibbs, queste saranno trattate nel prossimo capitolo. Questi modelli sono stati formalizzati proprio per sopperire alla mancanza di modelli che includessero nella soluzione l'effetto dell'errore legato all'attributo degli individui e riuscissero ad apprendere strutture a blocchi variegate e complesse, tipiche delle reti di organizzazioni criminali.

Per migliorare il processo d'apprendimento delle gerarchie interne a *La Lombardia*, ci limitiamo allo studio di soli 84 sospettati, coloro che provengono dalle 5 *locali* più numerose, mentre lasciamo fuori i 34 individui che appartengono a locali composte da meno di 7 affiliati. Questa scelta è utile anche per ridurre eventuali complicazioni che potrebbero venire fuori dalla non completa identificazione delle *locali* più piccole durante la fase di investigazione.

1.2 Analisi descrittiva dei dati

È naturale utilizzare l'immagine di una rete per catturare la disposizione degli elementi in un sistema e le loro interconnessioni. In un contesto matematico, come il nostro, questo vocabolo viene utilizzato in maniera equivalente al termine grafo, che descrive il tipo di rappresentazione impiegata per dati come la matrice di cui disponiamo.

Un grafo \mathcal{G} è dunque una struttura matematica costituita da due componenti: un insieme W di nodi e un insieme \mathcal{E} di legami. Gli elementi di \mathcal{E} sono coppie non ordinate (u,v) di nodi distinti, dove $u, v \in W$. Il numero di nodi $n = |W|$ e il numero di legami $M = |\mathcal{E}|$ sono denominati rispettivamente ordine e dimensione. Ci sono diversi tipi di grafo, noi ci concentreremo solo sul caso di grafo non orientato, ovvero quando non c'è alcun ordinamento nei nodi che definiscono un legame. Un modo per caratterizzare un grafo può essere attraverso la matrice di adiacenza \mathbf{Y} , definita come segue:

$$Y_{uv} = \begin{cases} 1 & \text{se } u, v \in \mathcal{E}, \\ 0 & \text{altrimenti.} \end{cases}$$

In altre parole \mathbf{Y} assume valori 1 per le entrate in cui l'indice riga-colonna corrisponde a nodi connessi da un legame e 0 per quelli che non lo sono. Tale matrice è simmetrica nel caso di grafi non orientati.

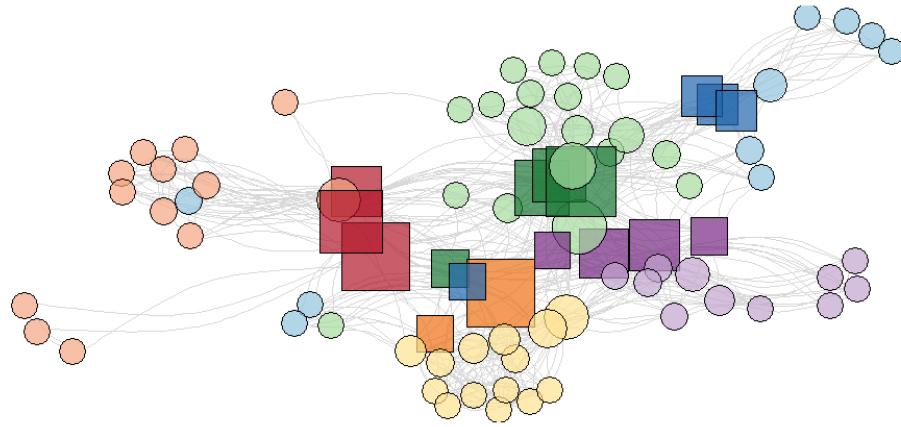


Figura 1.1: Rete *Infinito*

I nostri dati sono stati trasformati, come descritto in precedenza, in modo da assumere la forma di una matrice di adiacenza. Questi sono rappresentati da un grafo non orientato costituito da $n = 84$ nodi (affiliati) e $M = 663$ legami (co-partecipazione da parte di due affiliati ad uno stesso summit).

Per capire meglio la loro struttura andiamo a rappresentarli tramite il grafico in Figura 1.1, tale raffigurazione evidenzia il ruolo all'interno dell'organizzazione criminale di ciascun affiliato e la *locale* d'appartenenza desunta dagli atti, attraverso rispettivamente le forme e i colori. La posizione dei nodi è stata ottenuta attraverso un collocamento forzato e orientato (Fruchterman & Reingold, 1991). La dimensione dei nodi invece, è proporzionale alla *betweenness* corrispondente, misura di centralità che andremo a descrivere in seguito. I colori indicano la *locale* d'appartenenza tratta dagli atti giudiziari e la forma il ruolo all'interno dell'organizzazione, il quadrato si riferisce ai boss, mentre il cerchio agli affiliati. Attraverso tale rappresentazione, possiamo notare che la *locale* più numerosa è quella indicata dalla lettera C nel dataset e rappresentata dal colore verde nel grafico, questa infatti è composta da 19 affiliati 4 boss. Seguono la *locale* E descritta dal colore giallo con 14 affiliati e 2 boss e la A, raffigurata in rosso e costituita da 13 affiliati e 3 boss. Le due *locali* più piccole sono invece la D, viola, con 11 affiliati e 4 boss e la B, di colore azzurro, con 10 affiliati e 4 boss.

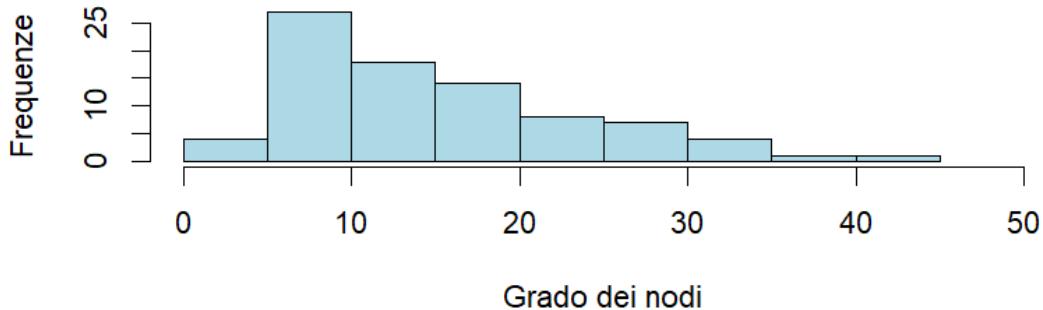


Figura 1.2: Istogramma delle frequenze del grado dei vertici

Andiamo adesso a descrivere la nostra rete di dati, per avere una prima idea della struttura sottostante all'organizzazione criminale. Si tratta di un grafo semplice, ovvero non presenta loop o legami multipli, e connesso, ossia ogni vertice è “raggiungibile” da tutti gli altri. Un nodo v è raggiungibile da un altro nodo u se esiste un percorso da u a v , dove un percorso è la forma più generale di connessione tra due nodi, descritta da una qualsiasi sequenza di nodi e legami che parte da v e arriva ad u .

Un modo per caratterizzare i nodi è attraverso il loro grado, il grado di un nodo v , d_v , è il numero di legami in \mathcal{E} incidenti a v , ossia il totale dei legami che hanno v come punto finale. Questa informazione ci permette di vedere quali nodi sono più connessi. Abbiamo evidenziato che il nodo con il grado maggiore, $d_v = 42$, è il nodo 62 che rappresenta un boss della *locale C*. È possibile raffigurare la distribuzione del grado attraverso l'istogramma delle frequenze del grado dei nodi $\{f_d\}_{d \geq 0}$, rappresentata in Figura (1.2) Notiamo come la maggior parte dei nodi ha grado compreso tra 5 e 10, dopo di che per gradi crescenti la frequenza diminuisce, indicando che sono pochi i nodi altamente connessi, in particolare i nodi con grado maggiore sono riferiti ai boss di *locali* differenti.

Un altro modo per analizzare l'importanza di un nodo all'interno della rete, è attraverso le misure di centralità. Esistono tre principali misure di centralità dette: *closeness*, *betweenness* e *eigenvector*. La *closeness* caratterizza la centralità di un nodo attraverso la sua vicinanza a tanti altri nodi:

$$C_{cl}(v) = \frac{1}{\sum_{u \in V} dist(v, u)},$$

dove $dist(v, u)$ è la distanza geodetica tra i nodi $u, v \in W$. La centralità varia inversamente a una misura della distanza totale tra un nodo e gli altri. L'indice di *betweenness*, precedente utilizzato per caratterizzare la dimensione dei nodi nella Figura 1.1, riassume la misura in cui un nodo è collocato tra altre coppie di nodi. L'importanza è dunque relativa a dove un nodo è collocato rispetto ai cammini del grafo. Un cammino è un percorso senza nodi o legami ripetuti. I nodi che giacciono su più cammini, sono i più cruciali all'interno della rete criminale.

$$C_B(v) = \sum_{s \neq t \neq v \in V} \frac{\sigma(s, t | v)}{\sigma(s, t)},$$

dove $\sigma(s, t | v)$ indica il numero totale di cammini più corti tra s e t che passano attraverso v e $\sigma(s, t)$ il numero totale di cammini più corti tra s e t , senza considerare se passano da v oppure no. L'indice *eigenvector* cattura l'idea che più i vicini di un nodo sono centrali, più il nodo stesso è centrale:

$$\mathbf{C}_{Ei} = \alpha \sum_{\{u, v\} \in E} C_{Ei}(u).$$

Il vettore $\mathbf{C}_{Ei} = (C_{Ei}(1), \dots, C_{Ei}(N_v))^T$ scaturisce dalla soluzione del problema dell'autovalore $\mathbf{Y}\mathbf{C}_{Ei} = \alpha^{-1}\mathbf{C}_{Ei}$, dove \mathbf{Y} è la matrice di adiacenza. Assumiamo che α^{-1} sia il più grande autovalore di \mathbf{Y} e \mathbf{Y}_{Ei} l'autovettore corrispondente. Quando il grafo è non orientato e connesso, come nel nostro caso, il più grande autovalore di \mathbf{Y} sarà semplice e il suo autovettore avrà tutte entrate diverse da zero e con lo stesso segno. Per convenzione si riporta il valore assoluto delle entrate che automaticamente si troverà tra 0 e 1, per l'ortonormalità dei vettori. La centralità *eigenvector* così definita fa sì che la centralità di ogni nodo dipenda sia dal numero che dalla qualità delle sue connessioni, infatti avere un grande numero di connessione è rilevante, allo stesso tempo se un nodo ha un minor numero di legami ma di maggiore qualità, potrebbe superare un nodo con più connessioni ma meno importanti. Per meglio interpretare queste misure di centralità dei nodi, è possibile rappresentarne i valori attraverso una configurazione circolare, dove i nodi più importanti, secondo ciascuna delle definizioni di centralità sopra descritte, saranno più vicini al centro. Possiamo osservare il grado e la *betweenness* nella Figura 1.3, mentre la *closeness* e *eigenvalue* nella Figura 1.4. In queste rappresentazioni abbiamo mantenuto i colori utilizzati nella Figura 1.1 per indicare la *locale* di appartenenza di ogni nodo/affiliato, le tonalità più scure indicano che tale individuo è stato identificato come boss. Analizzando le

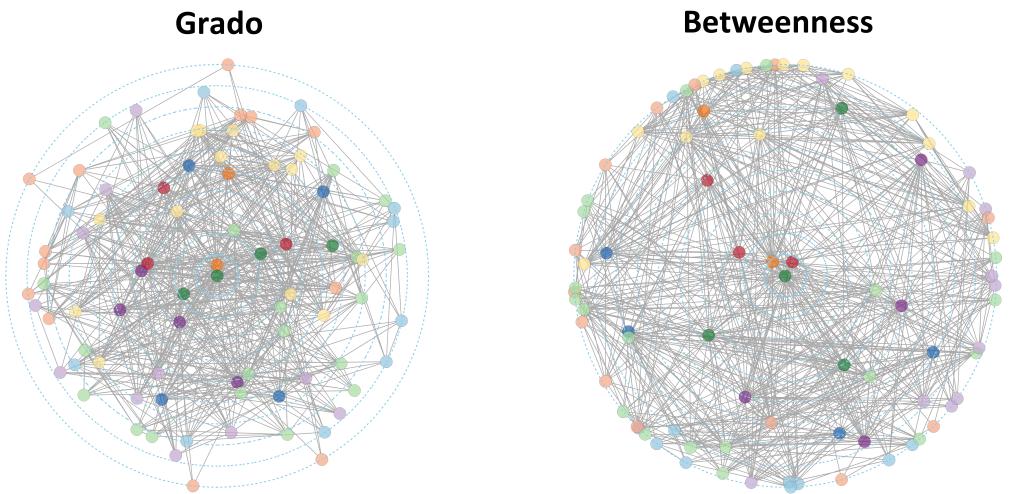


Figura 1.3: Grado e *betweenness*

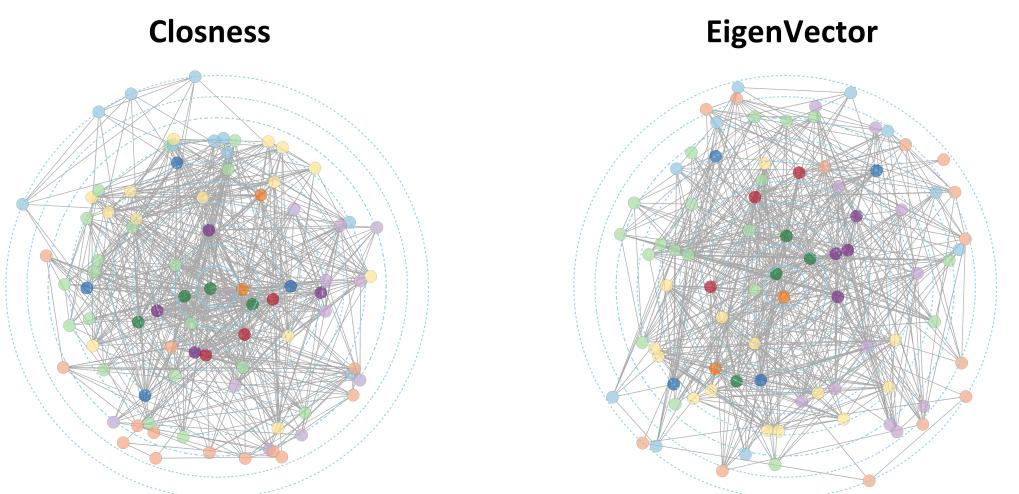


Figura 1.4: *closeness* e *eigenvalue*

Tabella 1.1: Cliques

Dimensione	1	2	3	4	5	6	7	8
N° cliques	84	663	2557	6248	10983	15070	16980	16069
Dimensione	9	10	11	12	13	14	15	16
N° cliques	12767	8382	4432	1825	560	120	16	1

Tabella 1.2: Cliques Massimi

Dimensione	4	6	7	8	9	10	11	12	16
N° cliques	2	2	13	9	10	3	5	5	1

diverse raffigurazioni, possiamo notare come il concetto di centralità si esprima in maniera diversa a seconda dell'indice utilizzato. Per quanto riguarda il grado dei nodi, ritroviamo quanto descritto in precedenza, i nodi più centrali sono due ed entrambi fanno riferimento a due boss. Gli stessi li possiamo ritrovare in posizione centrale anche nel grafico della *betweenness*, insieme a un terzo nodo sempre identificato da un boss. Il nodo più importante secondo l'indice *closeness* invece, coincide con quello che ha il grado più alto, mentre secondo l'*eigenvalue* corrisponde al terzo nodo trovato dall'indice *betweenness*. Tutte queste misure concordano nell'individuazione dei nodi più centrali, vale a dire più importanti all'interno della rete criminale, in tre nodi corrispondenti a tre boss appartenenti rispettivamente alle *locali* C (verde), E (arancione) ed A (rosso).

Passiamo adesso ad analizzare la coesione della nostra rete, ovvero la misura in cui sottoinsiemi di vertici sono coesi rispetto alla relazione che definisce i legami nella rete. Un modo per poter definire la coesione è attraverso la specificazione di un sotto-grafo. Un clique è un sotto-grafo completo, è quindi composto da un sottoinsieme di vertici tutti connessi tra di loro da un legame. Fare un censimento dei cliques di tutte le possibili dimensioni fornisce un'istantanea sulla struttura del grafo. Troviamo un riassunto di tutti i cliques rilevati nella nostra rete nella Tabella 1.1. Possiamo notare come i cliques rilevati arrivano ad assumere grandi dimensioni, in particolare il più grande è di dimensione 16, questo è sintomo di un'elevata connessione tra i nodi della rete. Dobbiamo ricordarci che spesso i cliques più grandi includono quelli più piccoli, è quindi possibile definire il clique massimo, ovvero un clique che non è un sottoinsieme di un clique più grande, questi sono riportati nella Tabella 1.2. Esistono nozioni di cliques più deboli, come k-nuclei: un sotto-grafo i cui nodi hanno come grado almeno k e tali che nessun altro sotto-grafo che rispetta la stessa condizione li contenga. Possiamo vedere la rappresentazione dei diversi nuclei nella Figura 1.5.

Scomposizione K-nuclei

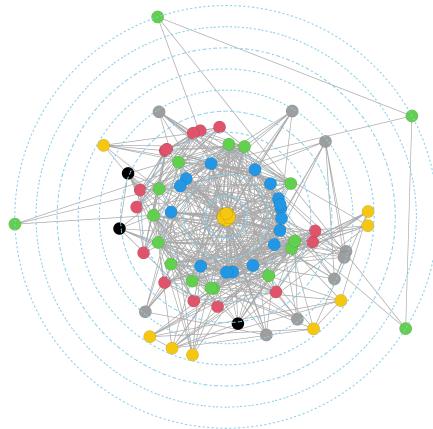


Figura 1.5: k-nuclei

Attraverso tale raffigurazione che segue uno schema circolare è possibile rilevare la scomposizione della rete in diversi strati. Altre informazioni riguardanti la coesione di una rete possono essere trovate applicando il concetto di frequenza relativa. Per esempio la densità di un grafo è definita come la frequenza di legami effettivamente realizzati rispetto a quelli potenziali:

$$den(\mathcal{G}) = \frac{2|\mathcal{E}|}{|W|(|W|-1)}.$$

Nel caso in cui \mathcal{G} fosse un sotto-grafico, la densità misurerebbe la sua vicinanza ad essere un clique. Applichiamo tale definizione alla nostra rete e osserviamo che questa ha densità pari a 0.19. Se prendiamo poi in considerazione i tre nodi più centrali trovati in precedenza e i rispettivi intorni, possiamo vedere come questi siano più densi dell'intera rete, hanno infatti valore di densità rispettivamente 0.431 per il nodo 40 che identifica un boss della *locale A*, 0.425 per il nodo 41 relativo ad un boss della *locale E* e 0.375 per il nodo 62 che descrive un boss della *locale C*. Il concetto di frequenza relativa viene anche utilizzato per definire la nozione di clustering in un grafo, di solito si fa riferimento al *coefficiente di clustering*, definito come segue:

$$cl_T(\mathcal{G}) = \frac{3\tau_\Delta(\mathcal{G})}{\tau_3(\mathcal{G})},$$

dove $\tau_\Delta(G)$ è il numero di triangoli nel grafo, per triangolo si intende un insieme di tre nodi tutti collegati tra di loro e $\tau_3(\mathcal{G})$ è il numero di triplette nel grafo, una tripletta è l'insieme di tre nodi che possono essere sia chiusi (tutti i nodi sono collegati tra di

loro) che aperti (esiste un nodo centrale collegato agli altri due nodi non connessi tra di loro, nel grafo). Questa misura è anche chiamata *transitività*, in quanto si riferisce alla frazione di triplette transitive. Nel nostro grafo circa il 60% delle triplette sono transitive.

Capitolo 2

Modelli probabilistici

In questo capitolo andremo ad illustrare tutti i modelli sottostanti lo sviluppo dei modelli per blocchi stocastici estesi, suddividendoli in due sezioni: quelli che ricorrono all’approccio bayesiano non parametrico e quelli specifici per dati di rete.

Inizieremo dunque la prima sezione introducendo il metodo bayesiano, descrivendo in particolare i metodi Markov Chain Monte Carlo, per poi concentrarci sul contesto non parametrico. In seguito descriveremo i modelli di mistura, in quanto costituiscono il fondamento dei modelli per blocchi stocastici. Successivamente tratteremo l’approccio non parametrico, in particolare la classe delle distribuzioni a priori di tipo Gibbs, descritte da [De Blasi et al. \(2015\)](#), utilizzate nei ESBM e descriveremo la distribuzione Stirling Gamma, sviluppata da [Zito et al. \(2024+\)](#), da utilizzare come distribuzioni a priori del Processo di Dirichlet, per ottenere risultati più robusti. Infine mostreremo l’algoritmo SALSO ([Dahl et al., 2022](#)), impiegato per trovare la partizione ottima al posto dell’algoritmo “avid” di [\(Wade & Ghahramani, 2018\)](#). Nella seconda parte descriveremo i modelli specifici per dati di rete, in particolare i modelli per blocchi stocastici (SBM), trattati da [Nowicki & Snijders \(2001\)](#) e la loro estensione (ESBM) ([Legramanti et al., 2022](#)).

2.1 Concetti preliminari

Prima di addentrarci nella descrizione dei modelli che ricorrono all’approccio bayesiano utilizzati nella costruzione dei modelli per blocchi stocastici estesi, andiamo a fare una breve introduzione a questo metodo, illustrando successivamente la sua evoluzione nell’ambito non parametrico.

Nell’approccio classico o frequentista l’idea fondamentale è che l’informazione sia data da una n-upla campionaria \mathbf{X} di realizzazioni, a cui si aggiunge un principio: il campionamento ripetuto. Le probabilità descrivono dunque le frequenze relative

di variabili casuali su un numero infinito di esperimenti. I parametri del modello sono considerati quantità fisse non note. Un modello statistico è caratterizzato da un insieme di distribuzioni $p(\mathbf{X} | \theta)$, dove i parametri θ assumono valori nello spazio parametrico. Tale distribuzione è detta verosimiglianza quando viene considerata come funzione di θ .

Nell'approccio bayesiano oltre all'informazione sperimentale, data dall'n-upla campionaria \mathbf{X} , si aggiunge un'informazione extra sperimentale, ovvero a disposizione prima dell'esperimento e il principio della verosimiglianza. Le probabilità quindi, oltre a descrivere le variabili casuali, vengono utilizzate per quantificare la conoscenza che si ha a priori sui parametri. A parità di evidenza sperimentale infatti, le informazioni sui parametri θ possono differire. Questi vengono quindi considerati come delle variabili casuali, dove la casualità riflette la nostra mancanza di conoscenza su di essi. A tale scopo, ai parametri viene assegnata una distribuzione di probabilità a priori $p(\theta)$ detta *prior*, la quale rappresenta l'informazione che si ha sui parametri prima di osservare l'n-upla campionaria. Spesso è conveniente specificare la distribuzione a priori attraverso una famiglia parametrica di distribuzioni. I parametri della *prior* sono detti iperparametri, questi possono essere fissi, oppure è possibile assegnare loro un'ulteriore distribuzione a priori detta *iperprior*, in tal caso si parla di modello bayesiano gerarchico. Una volta che la distribuzione a priori e la verosimiglianza sono state definite il modello è completamente specificato. L'idea cardine è prendere l'informazione sperimentale $p(\mathbf{X} | \theta)$ e l'informazione extra sperimentale $p(\theta)$ per ricavare nuova conoscenza su θ , aggiornata dall'esito sperimentale: $p(\theta | \mathbf{X})$. Questa distribuzione condizionata dei parametri è nota come distribuzione a posteriori ed è definita come segue:

$$\begin{aligned} p(\theta | \mathbf{X}) &= \frac{p(\mathbf{X} | \theta)p(\theta)}{\int p(\mathbf{X} | \theta)p(\theta)d\theta} \\ &= \frac{\text{verosimiglianza} \times \text{prior}}{\text{distribuzione marginale}}. \end{aligned}$$

L'obiettivo dunque, non è semplicemente la stima del parametro, ma la stima della distribuzione a posteriori che quantifica il nostro stato di conoscenza sul parametro dopo aver osservato i dati.

La modellazione dei dati attraverso l'approccio bayesiano può essere divisa in tre fasi:

- Distribuzione congiunta: il primo step coinvolge la formulazione del modello probabilistico, ovvero la distribuzione congiunta dei dati e del parametro, attraverso la specificazione della verosimiglianza e della distribuzione a priori.
- Inferenza: in seguito viene derivata la distribuzione a posteriori del parametro, spesso attraverso l'impiego di metodi d'approssimazione numerica come il campionamento Monte Carlo.
- Valutazione delle implicazioni: infine si controlla quanto bene il modello descriva i dati e si valutano le implicazioni della distribuzione a posteriori, calcolando le quantità d'interesse e prendendo decisioni.

2.1.1 Metodi Markov Chain Monte Carlo

Entriamo nel dettaglio della seconda fase (Inferenza). Spesso la distribuzione a posteriori del parametro $p(\theta | \mathbf{X})$ non è disponibile in forma chiusa, per risolvere questo inconveniente è possibile ricorrere a dei metodi simulativi, in particolare il metodo che verrà impiegato nei modelli per blocchi stocastici estesi è il campionamento Gibbs, il quale ricorre al metodo Monte Carlo applicato alle Catene di Markov. L'idea iniziale consiste nel creare una catena di Markov omogenea con distribuzione stazionaria propria pari a $p(\theta | \mathbf{X})$. Prima di descrivere questo processo simulativo, andiamo ad introdurre gli elementi che lo caratterizzano, denotiamo con $\{\mathbf{X}^n, n \in \mathbb{N}\}$ un processo stocastico a tempo discreto dotato della proprietà di markovianità:

$$P(\mathbf{X}^{n+1} \in \mathcal{A} | \mathbf{X}^n = \mathbf{x}^n, \dots, \mathbf{X}^0 = \mathbf{x}^0) = P(\mathbf{X}^{n+1} \in \mathcal{A} | \mathbf{X}^n = \mathbf{x}^n) \quad \forall \mathcal{A} \in \mathcal{S},$$

dove \mathcal{A} è un sottoinsieme misurabile dello spazio degli stati \mathcal{S} . Con $P(\mathbf{X}^{n+1} \in \mathcal{A} | \mathbf{X}^n = \mathbf{x}^n) = P(\mathbf{x}, \mathcal{A})$, andiamo invece ad indicare un kernel di transizione, ovvero una legge che consente l'aggiornamento ad un passo successivo, nel caso di kernel di transizione ad m passi utilizziamo la notazione: $P^m(\mathbf{x}, \mathcal{A})$. Infine con $p(\cdot)$ facciamo riferimento alla legge di distribuzione marginale del primo elemento della successione \mathbf{X}^0 , questa distribuzione è stazionaria se è la legge di distribuzione marginale di ogni elemento del processo. Ricorriamo al seguente teorema per poter applicare il metodo Monte Carlo a delle determinazioni che non sono indipendenti e identicamente distribuite, ma sono traiettorie di particolari catene. Sia $\{\mathbf{X}^n\}$ una catena di Markov omogenea, irriducibile e ricorrente, con kernel di transizione P e distribuzione stazionaria $p(\cdot)$, allora:

- $p(\cdot)$ è l'unica distribuzione stazionaria;

- se g è una funzione a valori reali, ovvero $E^p[|g(\mathbf{X})|] < \infty$, allora:

$$P(\bar{g}_n(\mathbf{X}) \xrightarrow{n \rightarrow +\infty} E[g(\mathbf{X})]) = 1,$$

dove $\bar{g}_n(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n g(\mathbf{X}_i)$;

- se la catena è anche aperiodica, allora:

$$\|P^n(\mathbf{X}, \cdot) - p(\cdot)\| \xrightarrow{n \rightarrow +\infty} 0 \quad \forall \mathbf{X} \in \mathcal{S}.$$

Nel caso in cui la distribuzione a posteriori d'interesse è $p(\theta \mid \mathbf{X})$, avremo una traiettoria per θ : $\{\theta^0, \dots, \theta^m, \theta^{m+1}, \dots, \theta^{m+n}, \dots\}$. Supponiamo che da m in poi sia valido il terzo punto del teorema, allora da quel momento possiamo pensare di avere delle determinazioni dalla distribuzione a posteriori $p(\theta \mid \mathbf{X})$, la prima parte della traiettoria, da θ^0 a θ^m , è detta periodo di *burn – in* e non verrà utilizzata. Le determinazioni da θ^{m+1} a θ^{m+n} , che abbiamo detto provenire dalla distribuzione d'interesse, le utilizzeremo come se fossero un campione Monte Carlo, per fare ciò n deve essere sufficientemente grande da verificare il secondo punto del teorema. Per quanto riguarda il punto di partenza, $p(\cdot)$, questo può essere scelto arbitrariamente se la catena è ben costruita e la vera legge iniziale è la distribuzione a posteri. Se θ è scomponibile: $\theta = (\theta_1, \dots, \theta_d)$, con $d > 1$ e siamo in grado di estrarre determinazioni casuali indipendenti e identicamente distribuite da opportune leggi di distribuzioni che descrivono ciascun elemento contenuto nella scomposizione di θ , possiamo ricorrere al campionamento Gibbs per generare la catena. Le distribuzioni a cui facevamo riferimento sono le seguenti distribuzioni a posteriori: $p(\theta_i \mid \theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_d, \mathbf{X})$ per $i = 1, \dots, d$, dette *full conditionals*. L'implementazione del campionamento Gibbs segue il seguente pseudo-codice:

- inizializzazione:* assegnare un primo valore a tutta la traiettoria $\theta^0 = (\theta_1^0, \dots, \theta_d^0)$
- aggiornamento:* per $i = 1, \dots, m + n$ genero θ^i dalle determinazioni delle “full conditionals”

$$\theta_1^i \text{ da } p(\theta_1 \mid \theta_2^{i-1}, \dots, \theta_d^{i-1}, \mathbf{X})$$

$$\theta_2^i \text{ da } p(\theta_2 \mid \theta_1^i, \theta_3^{i-1}, \dots, \theta_d^{i-1}, \mathbf{X})$$

.

.

$$\theta_d^i \text{ da } p(\theta_d \mid \theta_1^i, \theta_2^i, \dots, \theta_{d-1}^i, \mathbf{X})$$

2.1.2 Metodi bayesiani non parametrici

Prima di andare a vedere come si sviluppa l'approccio bayesiano in contesto non parametrico, andiamo a dare la definizione di modello non parametrico in generale: un modello statistico è detto non parametrico se la collezione di distribuzioni di probabilità da cui è caratterizzato non è indicizzata da un numero finito di parametri. L'approccio non parametrico alla statistica consiste dunque nel fare inferenza statistica senza ricorrere ad assunzioni parametriche per semplificare il problema. Come si evince dalla sua definizione, un modello non parametrico non è un modello senza parametri, ma un modello che non può essere caratterizzato da uno spazio parametrico finito dimensionale; possiamo pensare ad esso come ad un modello che possiede un numero infinito di parametri. Bisogna sempre ricordare che affermare di non fare assunzioni parametriche non equivale a non fare assunzioni. Possiamo adesso definire un modello non parametrico bayesiano, per fare ciò abbiamo bisogno di due elementi:

- $f \in \mathcal{F}$
- prior per f , definita su \mathcal{F} : $f \sim Q$,

dove \mathcal{F} è lo spazio delle distribuzioni non indicizzate da un parametro finito dimensionale e Q è una distribuzione su \mathcal{F} . L'esistenza della distribuzione a priori per Q è garantita dal teorema di rappresentazione di de Finetti: dato un vettore di variabili casuali X_1, \dots, X_n scambiabili, ovvero tali che $(X_1, \dots, X_n) \stackrel{d}{=} (X_{\pi(1)}, \dots, X_{\pi(n)})$ per ogni permutazione π di $\{1, \dots, n\}$, esiste Q su Θ che ci permette di scrivere la distribuzione di X_1, \dots, X_n in forma gerarchica come:

$$X_i \mid \tilde{P} \stackrel{iid}{\sim} \tilde{P}$$

$$\tilde{P} \sim Q,$$

dove \tilde{P} gioca il ruolo di distribuzione a priori non parametrica. Per definire \tilde{P} bisogna costruire una distribuzione aleatoria su di essa.

Successivamente andremo a descrivere i concetti chiave dell'inferenza bayesiana in un contesto scambiabile ed introdurremo le distribuzioni a priori non parametriche

discrete, concentrandoci su quelle di tipo Gibbs. Prima però, andiamo a trattare i modelli di mistura.

2.1.3 Modelli di mistura

Iniziamo la descrizione dei modelli bayesiani non parametrici con la trattazione dei modelli di mistura, in quanto costituiscono le fondamenta dei modelli per blocchi stocastici estesi, oggetto della nostra analisi.

I modelli di mistura possono essere utilizzati per problemi dove la popolazione è costituita da un certo numero di sotto popolazioni, ognuna descritta da un modello differente. Il principio base per la costruzione dei modelli di mistura consiste nell'introduzione del vettore \mathbf{z} , il quale specifica da quale componente proviene ciascuna osservazione, il vettore θ invece, descrive i parametri del modello che caratterizzano ciascuna osservazione. Un modello di mistura può essere visto come un modello gerarchico infatti, la variabile osservata \mathbf{X} viene modellata condizionatamente ai vettori \mathbf{z} e θ , che a loro volta sono dati da una distribuzione di probabilità. L'inferenza sui parametri (θ, \mathbf{z}) è data dalla distribuzione a posteriori congiunta. Supponiamo di voler modellare la distribuzione di $\mathbf{X} = (X_1, \dots, X_n)$ come una mistura di H componenti, assumendo di non conoscere da quale componente della mistura provenga ciascuna osservazione. Per $h = 1, \dots, H$ la distribuzione dell' h -esima componente $f_h(X_i | \theta_h)$, dipende dal vettore dei parametri θ_h . Il parametro che indica la proporzione di popolazione della componente h è π_h , con $\sum_{h=1}^H \pi_h = 1$. È solito assumere che le componenti della mistura provengano tutte dalla stessa famiglia, ma caratterizzata da parametri diversi. La distribuzione campionaria di X_i in questo caso diventa:

$$p(X_i | \theta, \pi) = \pi_1 f(X_i | \theta_1) + \pi_2 f(X_i | \theta_2) + \dots + \pi_H f(X_i | \theta_H).$$

La forma della distribuzione campionaria porta ad un naturale paragone con la classica struttura bayesiana appena introdotta. La distribuzione delle misture di probabilità $\pi = (\pi_1, \dots, \pi_H)$ può essere pensata come una distribuzione a priori discreta sui parametri θ_h , è più appropriato però pensare a questa distribuzione a priori, o mistura, come descrizione della variazione di θ sulla popolazione d'interesse. Adottando tale visione il modello di mistura ricorda più un modello gerarchico, questa

somiglianza aumenta introducendo l'indicatore non osservato delle variabili z_{ih}

$$z_{ih} = \begin{cases} 1 & \text{se l'unità } i\text{-esima proviene dalla componente } h, \\ 0 & \text{altrimenti.} \end{cases}$$

Dato π , la distribuzione di ogni vettore z_i è la seguente:

$$z_i = (z_{i1}, \dots, z_{iH}) | \pi \sim \text{Multinomiale}(1; \pi_1, \dots, \pi_H).$$

In questo caso i parametri della mistura λ sono pensati come iperparametri che determinano la distribuzione di \mathbf{z} . La distribuzione congiunta dei dati osservati \mathbf{X} e degli indicatori non osservati condizionatamente ai parametri del modello è la seguente:

$$p(\mathbf{X}, \mathbf{z} | \theta, \pi) = p(\mathbf{z} | \pi)p(\mathbf{X} | \mathbf{z}, \theta) = \prod_{i=1}^n \prod_{h=1}^H (\pi_h f(X_i | \theta_h))^{z_{ih}},$$

dove esiste un solo z_{ih} pari ad 1 per ogni i . Fino a questo punto abbiamo assunto noto e fisso il numero di componenti della mistura H , di solito questo però non è conosciuto.

La mistura finita è un caso speciale di una specificazione più generale:

$$p(X_i) = \int p(X_i | \theta) \pi(\theta) d\theta.$$

I modelli ad effetti casuali possono essere pensati come misture continue, nel senso che ogni X_i osservabile è una variabile casuale con distribuzione che dipende dai parametri θ_i . La distribuzione a priori dei parametri θ_i è data dalla distribuzione delle misture $\pi(\theta)$. L'approccio computazione per le misture continue segue quello delle misture finite.

I parametri di un modello non sono identificati se la stessa funzione di verosimiglianza è ottenuta da più di una scelta di parametri nel modello. Tutti i modelli di mistura non sono identificabili, ciò implica che la distribuzione non cambia se le etichette dei gruppi vengono permutate.

Una distribuzione a priori informativa ha l'effetto di identificare specifiche componenti caratteristiche di specifiche sotto popolazioni. La distribuzione a priori per i parametri del modello di mistura finito (θ, π) è il prodotto delle *prior*, tra loro indipendenti, di θ e π . Se il vettore degli indici della mistura $z_i = (z_{i1}, \dots, z_{iH})$ è modellato come una multinomiale con parametro π , allora la distribuzione a priori

coniugata naturale è la Dirichlet: $\pi \sim Dirichlet(\alpha_1, \dots, \alpha_H)$. La dimensione relativa dei parametri α_h descrive la media della distribuzione a priori per π ; e la somma degli α_h è una misura della forza della *prior*, tale somma indica anche la dimensione teorica del campione a priori. Utilizziamo θ per rappresentare il vettore contenente tutti i parametri delle componenti della mistura: $\theta = (\theta_1, \dots, \theta_H)$. Alcuni parametri potrebbero essere comuni a tutte le componenti, altri invece specifici per ogni singola componente. Non facciamo assunzioni sulla distribuzione a priori $p(\theta)$. Nelle misture continue, i parametri della distribuzione della mistura richiedono una distribuzione *iperprior*.

Per i modelli di mistura finiti spesso c'è incertezza riguardo il numero di componenti della mistura H da includere nel modello. Calcolare modelli per grandi valori di H , può essere molto dispendioso, perciò è consigliato iniziare con una mistura piccola e valutare via via l'adattamento del modello.

Le misture finite provengono dalla suposizione che ognuno degli $i = 1, \dots, n$ elementi nel campione appartenga ad una delle H sotto popolazioni, dove ogni classe latente possiede un valore diverso di uno o più parametri nel modello parametrico. Sia $z_i \in \{1, \dots, H\}$ l'indice della sotto popolazione per l'elemento i , questo indica lo stato della classe latente. Allora la riposta X_i per l'elemento i , condizionatamente a z_i segue la distribuzione

$$X_i | z_i \sim f(\theta_{z_i}, \phi),$$

dove $f(\theta, \phi)$ è una distribuzione parametrica con parametri ϕ che non variano attraverso le sotto popolazioni e i parametri θ_h che si riferiscono solo alla sotto popolazione h per $h = 1, \dots, H$. Supponendo che la proporzione di popolazione appartenente alla classe h sia $p(z_i = h) = \pi_h$, possiamo ottenere la verosimiglianza marginalizzando per lo stato latente della classe:

$$g(\mathbf{X} | \pi, \theta, \phi) = \sum_{h=1}^H \pi_h f(\mathbf{X} | \theta_h, \phi),$$

questa corrisponde ad una mistura finita con H componenti, dove alla componente h viene assegnato peso π_h . In generale g può essere approssimata da una varietà maggiore di verosimiglianze rispetto a f .

Possiamo vedere l'incorporazione delle sotto popolazioni latenti come una caratterizzazione realistica del vero meccanismo generatore dei dati e che tali sotto gruppi esistono davvero. In tal caso potrebbe essere interessante fare inferenza su tali sotto popolazioni e cercare di classificare gli individui secondo il gruppo d'appartenenza.

È possibile estendere i modelli di mistura finiti al caso non parametrico, attraverso i modelli di mistura bayesiani non parametrici. Questi possono essere definiti attraverso la rappresentazione gerarchica come segue:

$$X_i | \theta_i \stackrel{ind}{\sim} f(X | \theta_i) \quad i = 1, \dots, n$$

$$\theta_i | \tilde{P} \stackrel{iid}{\sim} \tilde{P} \quad i = 1, \dots, n$$

$$\tilde{P} \sim Q,$$

dove $\theta_1, \dots, \theta_n$ sono i parametri latenti, \tilde{P} è una misura di probabilità casuale discreta e Q rappresenta la sua distribuzione a priori. Una possibilità è utilizzare distribuzioni a priori di tipo Gibbs per Q , le quali verranno trattate nella sezione seguente. Il fatto che \tilde{P} sia discreta induce un clustering delle osservazioni generando legami tra i parametri latenti. Nello specifico, ci saranno $K_n = H$ valori distinti tra $\theta_1, \dots, \theta_n$ che partizionano le unità statistiche $\{1, \dots, n\}$ in H gruppi.

I modelli a blocchi stocastici, che utilizzeremo in seguito per l'analisi della *rete Infinito* sono una variante dei modelli di mistura bayesiani non parametrici.

2.1.4 Distribuzioni a priori non parametriche discrete: classe di tipo Gibbs

In questa sezione andremo in primo luogo ad introdurre le distribuzioni a priori non parametriche discrete, e in secondo luogo ci concentreremo sulla classe di distribuzioni a priori di tipo Gibbs (De Blasi et al., 2015), in quanto questa classe di distribuzioni a priori costituisce uno degli elementi di innovazione dell'estensione dei modelli per blocchi stocastici (ESBM), rispetto alla loro formulazione iniziale (SBM).

Supponiamo che $(X_n)_{n \geq 1}$ sia una sequenza infinita di osservazioni, dove ogni X_i assume valori in un insieme \mathbb{X} . Inoltre P_x è l'insieme di tutte le misure di probabilità su \mathbb{X} . Assumere $(X_n)_{n \geq 1}$ scambiabile equivale ad assumere l'esistenza di una distribuzione di probabilità Q su $P_{\mathbb{X}}$ tale che:

$$X_i | \tilde{P} \stackrel{iid}{\sim} \tilde{P}, \quad i = 1, \dots, n$$

$$\tilde{P} \sim Q, \tag{2.1}$$

per ogni $n \geq 1$. \tilde{P} è dunque un misura di probabilità casuale su \mathbb{X} e la sua distribuzione di probabilità Q , detta *misura di de Finetti*, rappresenta la distribuzione a priori, quando la (2.1) viene utilizzata come modello Bayesiano con un campione osservato

$X_i, i = 1, \dots, n$. Quando Q degenera su un sottospazio finito dimensionale di $P_{\mathbb{X}}$, si parla di inferenza parametrica, mentre quando il supporto di Q è infinito dimensionale, si tratta di problemi di inferenza non parametrica. Dato un campione X_1, \dots, X_n generato attraverso la (2.1), la distribuzione predittiva un passo avanti coincide con il valore atteso a posteriori di \tilde{P} :

$$P(X_{n+1} \in \cdot | X_1, \dots, X_n) = \int_{P_X} p(\cdot) Q(dp | X_1, \dots, X_n), \quad (2.2)$$

dove $Q(\cdot | X_1, \dots, X_n)$ indica la distribuzione a posteriori di \tilde{P} .

Prior discrete non parametriche, sono distribuzioni a priori che selezionano distribuzioni discrete con probabilità 1 e che giocano un ruolo chiave nelle procedure bayesiane non parametriche, queste caratteristiche sono condivise dalla classe di distribuzioni a priori di tipo Gibbs. Qualsiasi misura di probabilità casuale associata ad una distribuzione a priori discreta può essere infatti rappresentata come:

$$\tilde{P} = \sum_{j=1}^{\infty} \tilde{P}_j \delta_{L_j},$$

dove

- δ_c rappresenta la massa puntiforme unitaria concentrata in c ,
- $(\tilde{P}_j)_{j \geq 1}$ è una sequenza di variabili casuali non negative tali che $\sum_{j \geq 1} \tilde{P}_j = 1$ quasi certamente,
- $(L_j)_{j \geq 1}$ è una sequenza di variabili casuali che assumono valori in \mathbb{X} .

D'ora in avanti assumiamo inoltre che $(\tilde{P}_j)_{j \geq 1}$ e $(L_j)_{j \geq 1}$ siano indipendenti e che gli (L_j) siano indipendenti e identicamente distribuiti rispetto a una misura di probabilità diffusa, P^* su \mathbb{X} . Una sotto classe generale di misure di probabilità discrete casuali di questo tipo è detta *modelli per il campionamento di specie*, che analizzeremo a breve.

Quando si tratta delle X_i osservabili, la natura discreta di Q implica che qualsiasi campione X_1, \dots, X_n presenterà legami con probabilità positiva, generando perciò $K_n = H \leq n$ osservazioni distinte X_1^*, \dots, X_H^* con frequenze n_1, \dots, n_H e tali che $\sum_{h=1}^H n_h = n$. Quando si sceglie di analizzare strutture predittive specifiche la quantità principale da considerare, sia da un punto di vista teorico che matematico, è la probabilità di osservare un nuovo valore distinto non incluso nel campione

X_1, \dots, X_n , cioè:

$$P(X_{n+1} = "new" | X_1, \dots, X_n). \quad (2.3)$$

All'interno di questo contesto, le misure di probabilità casuali discrete possono essere caratterizzate in termini di partizioni casuali scambiabili. Dato che Q è discreta, \tilde{P} induce una partizione di X_1, \dots, X_n ben descritta dalla funzione di probabilità della partizione scambiabile (EPPF), data da:

$$p_H^n(n_1, \dots, n_H) = \int_{\mathbb{X}^H} E[\tilde{P}^{n_1}(dX_1) \dots \tilde{P}^{n_H}(dX_H)].$$

Questa indica la probabilità indotta da \tilde{P} di osservare un campione di dimensione n , X_1, \dots, X_n , che mostra $K_n = H$ osservazioni distinte con frequenze n_1, \dots, n_H o, equivalentemente, una specifica partizione in $K_n = H$ gruppi con frequenze n_1, \dots, n_H . Inoltre una EPPF soddisfa la regola d'addizione:

$$p_H^{(n)}(n_1, \dots, n_H) = p_{H+1}^{(n+1)}(n_1, \dots, n_H, 1) + \sum_{h=1}^H p_H^{(n+1)}(n_1, \dots, n_{h+1}, \dots, n_H).$$

L'identificazione della EPPF porta direttamente alla determinazione della distribuzione predittiva (2.2). Infatti, se X_1, \dots, X_n è un campione che contiene $H \leq n$ valori distinti con rispettive frequenze n_1, \dots, n_H , si ottiene:

$$P(X_{n+1} = "new" | X_1, \dots, X_n) = \frac{p_{H+1}^{(n+1)}(n_1, \dots, n_H, 1)}{p_H^{(n+1)}(n_1, \dots, n_H)},$$

e la distribuzione predittiva (2.2) è una combinazione lineare di $P^*(\cdot) = E(\tilde{P})$, che può essere interpretata come la combinazione di una distribuzione a priori ipotetica della forma di \tilde{P} e di una misura pesata delle osservazioni, ossia:

$$\begin{aligned} P(X_{n+1} \in \cdot | X_1, \dots, X_n) &= \frac{p_{H+1}^{(n+1)}(n_1, \dots, n_H, 1)}{p_H^{(n+1)}(n_1, \dots, n_H)} P^*(\cdot) \\ &\quad + \sum_{h=1}^H \frac{p_{H+1}^{(n+1)}(n_1, \dots, n_{h+1}, \dots, n_H, 1)}{p_H^{(n+1)}(n_1, \dots, n_H)} \delta_{X_j}^*(\cdot). \end{aligned}$$

Possiamo notare che la regola d'addizione garantisce che la parte destra sommi ad 1 se valutata sull'intero spazio \mathbb{X} .

Le distribuzioni a priori non parametriche discrete sono adatte per rispondere alle difficoltà inferenziali che provengono da problemi di campionamento delle specie e nella modellazione delle misture. Andiamo dunque a trattare questi due contesti.

Consideriamo una misura di probabilità discreta, $\tilde{P} = \sum_{j=1}^{\infty} \tilde{P}_j \delta_{L_j}$, questa può essere vista come strumento per descrivere la struttura di una popolazione composta da diverse classi, ognuna presente con una certa proporzione. Tali proporzioni sono modellate attraverso la misura di probabilità discreta, come proporzioni casuali \tilde{P}_j . Per questo motivo, si possono utilizzare in modo equivalente gli L_i o gli interi positivi $\{1, 2, \dots\}$ per etichettare diverse classi che possono essere campionate. Infatti, se $(\xi_n)_{n \geq 1}$ è una sequenza di valori interi ausiliari, tale che $P(\xi_n = j | \tilde{P}) = \tilde{P}_j$, per ogni n e per ogni j , il modello (2.1) corrisponde ad assumere che $X_i = L_{\xi_i}$. Gli X_n possono dunque essere interpretati come le etichette delle classi osservate, poiché, a causa della natura diffusa di P^* , due punti X_i e X_j , per $i \neq j$, differiscono se e solo se anche ξ_i e ξ_j sono diversi tra loro. Inoltre, $P(\xi_i = \xi_j) > 0$ per ogni $i \neq j$, questo comporta che l'osservazione i-esima e j-esima possono riferirsi alla stessa classe con probabilità positiva. Oltre tutto, una sequenza scambiabile $(X_n)_{n \geq 1}$, per cui vale la (2.1), dove \tilde{P} è un modello di campionamento per specie, prende il nome di sequenza di campionamento per specie.

Le priori non parametriche discrete sono anche utilizzate nella costruzione di modelli di mistura gerarchici, questi di solito vengono utilizzati per la stima della densità e per il clustering. Consideriamo il caso di una stima di densità univariata, dove $f(\cdot | \cdot)$ indica un kernel definito su $\mathbb{R} \times \mathbb{X}$ che assume valori in \mathbb{R}^+ , tale che $\int_{\mathbb{R}} f(t | x) dy = 1$, per ogni x contenuta in \mathbb{X} . Le osservazioni vengono quindi da una sequenza $(T_n)_{n \geq 1}$ di variabili casuali reali tali che:

$$\begin{aligned} T_i | X_i &\stackrel{ind}{\sim} f(\cdot | X_i) \quad i = 1, \dots, n \\ X_i | \tilde{p} &\stackrel{iid}{\sim} \tilde{P} \quad i = 1, \dots, n \\ \tilde{P} &\sim Q. \end{aligned} \tag{2.4}$$

La scelta tipica per \tilde{P} è rappresentata dal Processo di Dirichlet, introdotto da Fruchterman & Reingold (1991), distribuzione a priori che analizzeremo in seguito in quanto appartenente alla classe di tipo Gibbs, il quale porta al modello di mistura Processo di Dirichlet.

Oltre che per la stima della densità, questi modelli sono utilizzati anche in contesti di clustering. Infatti, sia $(X_n)_{n \geq 1}$ una sequenza di elementi casuali scambiabili latenti e $K_n = H$, il numero non osservato di valori distinti tra X_1, \dots, X_n , e il numero di gruppi in cui T_1, \dots, T_n possono essere raggruppati. L'inferenza a posteriori per $K_n = H$ è di grande importanza, la specificazione di una distribuzione a priori di tipo

Gibbs per \tilde{P} nella (2.4), permette un'individuazione efficace del numero di gruppi generati dai dati.

Le distribuzioni a priori di tipo Gibbs sono un caso particolare di distribuzioni a priori non parametriche discrete, che possiedono proprietà importanti sia a livello teorico che applicativo:

- permettono di individuare soluzioni in tempo polinomiale a problemi legati a classi generali di distribuzioni a priori non parametriche, infatti consentono di dividere la regola di previsione in due step e di evidenziare le quantità chiave, permettendo semplificazioni delle espressioni rilevanti;
- ammettono una caratterizzazione predittiva intuitiva che giustifica il loro impiego come assunzione corretta nel processo di apprendimento;
- includono vari casi speciali come il Processo di Dirichlet (Fructerman & Reingold, 1991) e di Pitman-Yor (Pitman, 2006), ma non solo.

Queste distribuzioni a priori inoltre, forniscono una classificazione interessante ed utile dei modelli di campionamento per specie dal punto di vista della struttura della probabilità indotta nella generazione di un nuovo valore. Per raggiungere tale scopo la quantità chiave è: $P(X_{n+1} = "new" | X_1, \dots, X_n)$, che rappresenta la probabilità di generare un nuovo valore dato il passato, associato ad un modello di campionamento per specie. Secondo la sua struttura possiamo classificare le distribuzioni a priori sottostanti in tre categorie principali. Indichiamo con Θ un parametro finito dimensionale che possibilmente entri nella specificazione di \tilde{P} in (2.4). In generale abbiamo $P(X_{n+1} = "new" | X_1, \dots, X_n) = f(n, H, n_1, \dots, n_H, \Theta)$, che vuol dire che la probabilità di ottenere una nuova osservazione dipende dalla dimensione del campione n , dal numero di valori distinti H , dalla loro frequenza (n_1, \dots, n_H) e dal parametro Θ . Indichiamo $P(X_{n+1} = "new" | X_1, \dots, X_n)$ con $f(n, H, \Theta)$ se non dipende da (n_1, \dots, n_H) e con $f(n, \Theta)$ se non dipende né da (n_1, \dots, n_H) , né da H . Sia \tilde{P} un modello di campionamento per specie, allora vale la seguente classificazione dal punto di vista della struttura di probabilità di generare un nuovo valore:

- $P(x_{n+1} = "new" | x_1, \dots, x_n) = f(n, \Theta)$ se e solo se \tilde{P} è un processo di Dirichlet,
- $P(x_{n+1} = "new" | x_1, \dots, x_n) = f(n, H, \Theta)$ se e solo se \tilde{P} è di tipo Gibbs,
- $P(x_{n+1} = "new" | x_1, \dots, x_n) = f(n, H, n_1, \dots, n_k, \Theta)$ altrimenti.

Nonostante il Processo di Dirichlet abbia dimostrato di funzionare bene in svariati contesti applicativi, da un punto di vista concettuale sembra troppo restrittivo lasciare che la probabilità di generare un nuovo valore dipenda solo dalla dimensione del campione n e dal parametro di massa Θ e riassumere tutte le altre informazioni contenute nei dati attraverso una specificazione adeguata del parametro Θ . Sarebbe desiderabile che tale probabilità dipendesse esplicitamente anche dal numero di valori distinti osservati H , poiché questo valore riassume l'eterogeneità nel campione, per questo motivo ricorriamo alle distribuzioni a priori di tipo Gibbs. A seconda della situazione specifica, si potrebbe voler modellare la $P(X_{n+1} = \text{"new"} | X_1, \dots, X_n)$ come una funzione crescente o decrescente di H , questo corrisponderà a una distribuzione a priori di tipo Gibbs con un parametro, rispettivamente positivo o negativo. Il terzo caso, che corrisponde alla situazione più generale e dove la previsione di un nuovo valore dipende esplicitamente da tutte le informazioni fornite dai dati, sarebbe la struttura di previsione maggiormente desiderabile. Nonostante ciò ci sono due problemi operativi particolarmente significativi che devono essere presi in considerazione. Da una parte il caso generale causa seri ostacoli analitici e le distribuzioni a priori devono quindi essere studiate caso per caso, portando a espressioni complesse. Dall'altra, non è chiaro come andrebbe specificata in modo esplicito la dipendenza della probabilità di osservare una nuova classe dalle frequenze osservate n_1, \dots, n_H , così da riflettere un'opinione sul meccanismo di previsione dei dati. Sarebbe ragionevole che tale opinione sulla distribuzione a priori fosse codificata attraverso il parametro finito dimensionale Θ . La classificazione di cui sopra mostra nettamente come le distribuzioni a priori di tipo Gibbs permettano di risolvere problemi matematici in tempo polinomiale, grazie ad una precisa assunzione semplificatrice sulla struttura di previsione. Tutto sommato una tale assunzione sembra essere un compromesso soddisfacente tra generalità e semplicità analitica.

Dopo aver fornito una caratterizzazione predittiva delle distribuzioni a priori di tipo Gibbs, andiamo adesso a dare una definizione equivalente, più utile per analizzare le proprietà delle distribuzioni. Una distribuzione a priori non parametrica discreta \tilde{P} associata ad uno schema scambiabile del tipo (2.1), può essere caratterizzata in termini della funzione di probabilità della partizione scambiabile associata $\{p_H^{(n)} : n \geq 1, 1 \leq H \leq n\}$, definita come in (2.3). Di conseguenza, si definisce una distribuzione a priori di tipo Gibbs come un modello di campionamento per specie, per $a > 0$ che indica il fattoriale crescente con $(a)_n = a(a+1)\dots(a+n-1)$

per ogni $n \geq 1$ e posto $(a)_0 = 1$, tale che:

$$p_H^{(n)}(n_1, \dots, n_H) = V_{n,H} \prod_{h=1}^H (1 - \sigma)_{n_h-1}, \quad (2.5)$$

per ogni $n \geq 1$, $H \leq n$ e interi positivi n_1, \dots, n_H tali che $\sum_{h=1}^H n_h = n$, dove $\sigma < 1$ e l'insieme di pesi non negativi $\{V_{n,H} : n \geq 1, 1 \leq H \leq n\}$ soddisfa la seguente equazione ricorsiva:

$$V_{n,H} = (n - \sigma H)V_{n+1,H} + V_{n+1,H+1}$$

per ogni $H = 1, \dots, n$ e $n \geq 1$, con $V_{1,1} = 1$. Alla luce della (2.5) è possibile descrivere le ragioni della soluzione in tempi polinomiali delle distribuzioni a priori di tipo Gibbs in termini più matematici, infatti la forma per prodotto delle loro EPPF permettono di gestire in modo conveniente le frequenze n_h . Data la (2.5), la probabilità di ottenere una nuova osservazione condizionatamente al campione x_1, \dots, x_n tale che $K_n = H$ è:

$$P(X_{n+1} = \text{"new"} \mid X_1, \dots, X_n) = \frac{V_{n+1,H+1}}{V_{n,H}} = f(n, H, \Theta).$$

La definizione (2.5) implica che le distribuzioni predittive indotte siano:

$$P(X_{n+1} \in \cdot \mid X_1, \dots, X_n) = \frac{V_{n+1,H+1}}{V_{n,H}} P^*(.) + \frac{V_{n+1,H}}{V_{n,H}} \sum_{h=1}^H (n_h - \sigma) \delta_{X_h}(.). \quad (2.6)$$

La distribuzione predittiva è dunque una combinazione lineare convessa della *prior* ipotizzata P^* con la forma di \tilde{P} e della distribuzione empirica pesata $\hat{P}_n = (n - H\sigma)^{-1} \sum_{h=1}^H (n_h - \sigma) \delta_{X_h}$. La struttura predittiva mostra alcune caratteristiche interessanti, in particolare il meccanismo di allocazione della massa predittiva tra dati "nuovi" e dati precedentemente osservati può essere diviso in due stadi. Dato un campione X_1, \dots, X_n , il primo step consiste nell'allocare la massa tra un nuovo valore X_{H+1}^* e un insieme di valori osservati $\{X_1^*, \dots, X_H^*\}$. Questo passo dipende solo da n e da H e non dalle frequenze n_1, \dots, n_H . Il secondo step è il seguente: condizionatamente al fatto che X_{n+1} sia un nuovo valore, questo è campionato dalla misura base P^* , mentre se X_{n+1} coincide con uno dei valori precedentemente osservati X_h^* , per $h = 1, \dots, H$, le probabilità corrispondenti sono determinate dalla dimensione n_h di ciascun gruppo e da σ . Perciò, anche se le frequenze n_h non influenzano la probabilità di allocare un valore previsto tra "vecchio" e "nuovo", sono esplicitamente coinvolte condizionatamente al valore previsto di

coincidere con uno precedentemente osservato: più un'osservazione passata viene individuata, maggiore sarà la probabilità di riosservarla.

Esiste anche una connessione stretta tra le distribuzioni a priori di tipo Gibbs, la partizione che queste inducono e i modelli di prodotti di partizioni scambiabili. Se Π_n rappresenta una partizione casuale di un insieme di interi $\{1, \dots, n\}$, un modello di prodotti di partizioni corrisponde alla distribuzione di probabilità per Π_n , rappresentata come segue:

$$P(\Pi_n = \{S_1, \dots, S_H\}) \propto \prod_{h=1}^H \rho(S_h), \quad (2.7)$$

dove $\rho(\cdot)$ è detta funzione di coesione. Sia $|S| = \text{cardinalità}(S)$ e imponiamo che la funzione di coesione $\rho(\cdot)$ dipenda solo dalla cardinalità dell'insieme S , ovvero $\rho(S_h) := \rho(|S_h|) = \rho(n_h)$, si tratta di una scelta ragionevole per una funzione di coesione. La partizione casuale è dunque, per ogni $n \geq H \geq 1$, la partizione casuale indotta da una sequenza scambiabile se e solo se $\rho(n_h) = (1 - \sigma)_{n_h-1}/n_h!$, per $h = 1, \dots, H$ e $\sigma \in [-\inf, 1]$, con la condizione che $(1 - \sigma)_{n_h-1} = 1$ quando $\sigma = -\inf$ e che Π_n si riduca a una partizione di un singolo elemento quando $\sigma = 1$. Questo è equivalente a dire che Π_n è di tipo Gibbs. Di conseguenza, le misure di probabilità casuali che inducono modelli di prodotti di partizioni scambiabili con funzione di coesione che dipende dalla cardinalità, come:

$$X_h^* \mid \Pi_n \stackrel{iid}{\sim} P^* \quad h = 1, \dots, H$$

$$\Pi_n \sim \text{distribuzione di prodotti di partizioni con } \rho(S) = \rho(|S|),$$

coincidono con la famiglia di distribuzioni a priori di tipo Gibbs. Tale connessione è utile per riuscire ad incorporare gli attributi specifici per nodo nei modelli per blocchi stocastici estesi, come vedremo in seguito.

Andiamo adesso a studiare alcune distribuzioni a priori che fanno parte di questa famiglia e che utilizzeremo in seguito nell'implementazione del modello: Dirichlet-Multinomiale, Processo di Dirichlet, Pitman-Yor e Gneden. Prima di entrare nel dettaglio di ognuna di esse andiamo a riscrivere la (2.6):

$$p(X_{n+1} = h \mid X_1, \dots, X_n) \propto \begin{cases} \frac{V_{n+1,H}}{V_{n,H}}(n_h - \sigma) & \text{per } h = 1, \dots, H, \\ \frac{V_{n+1,H+1}}{V_{n,H}} & \text{per } h = H + 1. \end{cases} \quad (2.8)$$

Iniziamo dalla distribuzione a priori Dirichlet-Multinomiale, questa è caratterizzata da tre iperparametri: $\sigma < 0$, $\beta = -\sigma$ e $\bar{H} \in \{1, 2, \dots\}$, quest'ultimo indica una conoscenza a priori sul numero di gruppi ed è fisso. Il peso $V_{n,H}$ è caratterizzato come segue:

$$V_{n,H} = \frac{\beta^{H-1}}{(\beta\bar{H} + 1)_{n-1}} \prod_{h=1}^{H-1} (\bar{H} - h) \mathbb{1}(H \leq \bar{H}),$$

dove n_h indica il numero di osservazioni presenti nel gruppo h . Conseguentemente la (2.8) diventa:

$$p(X_{n+1} = h \mid X_1, \dots, X_n) \propto \begin{cases} \frac{1}{\beta\bar{H}+n} (n_h + \beta) & \text{per } h = 1, \dots, H, \\ \frac{\beta}{\beta\bar{H}+n} (\bar{H} - h) \mathbb{1}(H \leq \bar{H}) & \text{per } h = H + 1. \end{cases}$$

E il numero atteso di gruppi è:

$$E(H \mid \bar{H}, \sigma) = \bar{H} - \bar{H} \frac{(\bar{H}|\sigma| - |\sigma|)_n}{(\bar{H}|\sigma|)_n}.$$

Continuiamo con la distribuzione a priori Processo di Dirichlet (Fruchterman & Reingold, 1991), questa è caratterizzata da un solo iperparametro $\alpha > 0$, σ è infatti posto pari a 0 e \bar{H} è infinito. Il peso $V_{n,H}$ assume dunque la seguente forma:

$$V_{n,H} = \frac{\alpha^H}{(\alpha)_n},$$

e la (2.8) diviene:

$$p(X_{n+1} = h \mid X_1, \dots, X_n) = \begin{cases} \frac{1}{\alpha+n} n_h & \text{per } h = 1, \dots, H, \\ \frac{\alpha}{\alpha+n} & \text{per } h = H + 1. \end{cases}$$

Mentre il numero atteso di gruppi è:

$$E(H \mid \alpha) = \sum_{i=1}^n \frac{\alpha}{\alpha + i - 1}.$$

La distribuzione a priori Pitman-Yor (Pitman, 2006) a differisce dalla precedente solo per la presenza di $\sigma \in [0, 1)$ e $\alpha > -\sigma$. Il peso $V_{n,H}$ diventa quindi

$$V_{n,H} = \prod_{h=1}^{H-1} \frac{\alpha + h\sigma}{(\alpha + 1)_{n-1}}$$

la (2.8) prende il seguente aspetto:

$$p(X_{n+1} = h \mid X_1, \dots, X_n) = \begin{cases} \frac{1}{\alpha+n}(n_h - \sigma) & \text{per } h = 1, \dots, H, \\ \frac{\alpha+\sigma H}{\sigma+n} & \text{per } h = H + 1. \end{cases}$$

Mentre il numero atteso di gruppi è:

$$E(H \mid \alpha, \sigma) = \frac{(\alpha + \sigma)_n}{\sigma(\alpha + 1)_{n-1}} - \frac{\alpha}{\sigma}.$$

Infine la distribuzione a priori di Gneden ([Gneden, 2010](#)) ha come iperparametro $\gamma \in (0, 1)$, σ è fisso e pari a $\sigma = -1$, mentre \bar{H} diventa casuale attraverso l'assegnazione di una *iperprior*: $p(\bar{H} = i) = \frac{\gamma(1-\gamma)_{h-1}}{h!}$, questa garantisce la consistenza della distribuzione a posteriori per la stima della struttura in gruppi. Il peso $V_{n,H}$ diviene:

$$V_{n,H} = \frac{(\gamma)_{n-H} \prod_{h=1}^{H-1} (h^2 - \gamma h)}{\prod_{i=1}^{n-1} (i^2 + \gamma i)},$$

di conseguenza la (2.8) assume la forma:

$$p(X_{n+1} = h \mid X_1, \dots, X_n) \propto \begin{cases} \frac{1}{\gamma+n-\bar{H}} (n^2 - \gamma n) & \text{per } h = 1, \dots, H, \\ \frac{H^2 - \gamma H}{n^2 - \gamma n} & \text{per } h = H + 1. \end{cases}$$

Per quanto riguarda il numero atteso di gruppi invece non esiste ancora una forma chiusa per calcolarlo.

Le distribuzioni a priori Dirichlet-Multinomiale, Processo di Dirichlet, Pitman-Yor e Gneden forniscono vari meccanismi generativi realistici per la strutturazione in gruppi delle reti criminali, permettendo a chi sta svolgendo le analisi di scegliere quale tra queste meglio si adatta al caso in esame, oppure consentendo di testare diverse specificazioni durante la fase di modellazione. Le dostribuzioni a priori Processo di Dirichlet e Pitman-Yor, per esempio, possono fornire costruzioni utili per l'analisi di organizzazioni criminali relativamente instabili e frammentate, come le reti terroristiche, le quali sono caratterizzate da piccoli nuclei multipli e anche "lupi solitari". Il crimine organizzato come la 'Ndrangheta invece è contraddistinto da un'architettura più stabile e fortemente regolata da moduli, la quale potrebbe supportare l'utilizzo di distribuzioni a priori con un numero finito di gruppi nella popolazione, come la Dirichlet-Multinomiale e la Gneden. Ovviamente nella maggior parte degli studi forensi tale numero, \bar{H} , non è noto e perciò quantificare l'incertezza

attraverso la distribuzione a priori di Gnedin fornisce una soluzione più realistica rispetto a fissare \bar{H} come nella Dirichlet-Multinomiale.

2.1.5 Stirling-Gamma

In questa sezione andiamo a descrivere la distribuzione Stirling-Gamma (Zito et al., 2024+), in quanto, se impiegata come distribuzione a priori del parametro α del Processo di Dirichlet, può portare a partizioni a posteriori più robuste, rispetto a quelle che otterremmo mantenendo fisso il valore di tale parametro.

La distribuzione a priori Processo di Dirichelet è particolarmente sensibile al parametro di precisione, infatti l'impiego di diversi valori di α può portare a distribuzioni a posteriori molto differenti di H , numero di gruppi, anche quando i gruppi presenti nei dati sono sufficientemente separati. Tale mancanza di robustezza è problematica quando la partizione a posteriori è l'oggetto d'interesse inferenziale, come nel caso di clustering o individuazione di una comunità, ovvero il contesto della nostra analisi. Randomizzare il parametro di precisione attraverso una distribuzione a priori $p(\alpha)$ però, può attenuare questo comportamento. La distribuzione a priori $p(\alpha)$ più comunemente adottata è la distribuzione gamma, un'altra valida alternativa è la distribuzione a priori di Jeffrey. Tuttavia, tali distribuzioni portano ad ottenere distribuzioni a priori sulla partizione analiticamente intrattabili, questo previene un'elicitazione trasparente degli iperparametri della distribuzione a priori e complica l'inclusione dell'informazione disponibile sul clustering dei dati. Inoltre è stato mostrato che la distribuzione di H proveniente da un Processo di Dirichlet può essere approssimata con una distribuzione di Poisson quando α è fisso, tale approssimazione non è disponibile nel caso di α casuale. Una prior $p(\alpha)$ che fa fronte a tali problematiche è detta Stirling-Gamma, a seguito della sua connessione con i numeri Stirling e la distribuzione gamma; sotto un opportuno riscalamento logaritmico infatti, la Stirling-Gamma è equivalente alla gamma in un caso limite.

Quando α segue una distribuzione a priori Stirling-Gamma, la partizione casuale si dice provenire da un processo Stirling-Gamma, questa appartiene ancora alla classe dei modelli di partizione di tipo Gibbs. La Stirling-Gamma ha l'ulteriore vantaggio fondamentale di essere la distribuzione a priori coniugata alla legge della partizione casuale del Processo di Dirichlet se uno dei suoi iperparametri è pari a n . Questo accade perché la partizione casuale di un modello di mistura Processo di Dirichlet appartiene alla classe delle famiglie esponenziali naturali, che ammette una distribuzione a priori coniugata.

Andiamo adesso a definire la distribuzione Stirling-Gamma e a descriverne le principali proprietà. Una variabile casuale positiva segue una distribuzione Stirling-Gamma con parametri $a, b > 0$ e $m \in \mathbb{N}$ che soddisfano $1 < a/b < m$, se la sua funzione di densità è:

$$p(\alpha) = \frac{1}{S_{a,b,m}} \frac{\alpha^{a-1}}{\{(\alpha)_m\}^b},$$

$$\text{dove } S_{a,b,m} = \int_{\mathbb{R}_+} \frac{\alpha^{a-1}}{\{(\alpha)_m\}^b},$$

e si indica con $\alpha \sim SG(a, b, m)$.

Il nome della distribuzione Stirling-Gamma proviene dalla presenza del fattoriale crescente nella funzione di densità, la cui espansione polinomiale definisce i numeri di Stirling del primo tipo e la sua stretta connessione con la distribuzione gamma. Infatti, se $\alpha \sim Sg(a, b, m)$, allora vale la seguente convergenza:

$$\alpha \log m \rightarrow \gamma, \quad \gamma \sim Ga(a - b, b), \quad m \rightarrow \infty,$$

dove $Ga(a_0, b_0)$ indica una distribuzione gamma di media a_0/b_0 e varianza a_0/b_0^2 . Questa convergenza ha due implicazioni principali, la prima è che la densità della Stirling-Gamma $Sg(a, b, m)$ assomiglia progressivamente a una $Ga(a - b, b \log m)$, all'aumentare di m . La seconda è che $\alpha \rightarrow 0$ in probabilità per $m \rightarrow \infty$, con un tasso di convergenza logaritmico, attraverso un'applicazione diretta del teorema di Slutsky. Rispetto alla distribuzione Gamma, la distribuzione Stirling-Gamma ha una coda destra più pesante. La funzione di densità di una Stirling-Gamma è propria, infatti $S_{a,b,m} < \infty$ se e solo se $1 < a/b < m$. È interessante notare come la costante di normalizzazione $S_{a,b,m}$ giochi un ruolo chiave nel calcolo dei momenti della distribuzione. Infatti, se $\alpha \sim Sg(a, b, m)$ e supponendo che $0 < s < mb - a$, allora:

$$E(\alpha^s) = \frac{S_{a+s,b,m}}{S_{a,b,m}}$$

Quando $s > mb - a$ invece, si ottiene $E(\alpha^s) = \infty$. In generale non è ancora disponibile un'espressione analitica.

Come abbiamo detto quando il parametro di precisione di un Processo di Dirichlet segue una distribuzione Stirling-Gamma $\alpha \sim Sg(a, b, m)$ abbiamo un processo Stirling-Gamma, appartenente alla famiglia di tipo Gibbs con $\sigma = 0$. Per questo motivo la funzione di probabilità della partizione scambiabile associata è direttamente

disponibile:

$$P(\Pi_n = \{S_1, \dots, S_H\}) = \frac{\mathcal{V}_{a,b,m}(n, H)}{\mathcal{V}_{a,b,m}(1, 1)} \prod_{h=1}^H (n_h - 1)!,$$

dove i coefficienti sono pari a

$$\mathcal{V}_{a,b,m}(n, H) = \int_{\mathbb{R}_+} \frac{\alpha^{a+H-1}}{\{(\alpha)_m\}^b (\alpha)_n} d\alpha.$$

È possibile notare come i coefficienti di tipo Gibbs siano $V_{n,H} = \mathcal{V}_{a,b,m}(n, H)/\mathcal{V}_{a,b,m}(1, 1)$, per $H = 1, \dots, n$ e $n \geq 1$, con $V_{1,1} = 1$ e che la ricorsività in avanti è soddisfatta dal momento che $\mathcal{V}_{a,b,m}(n, H) = n \mathcal{V}_{a,b,m}(n+1, H) + \mathcal{V}_{a,b,m}(n+1, H+1)$. Inoltre, essendo una distribuzione a priori di tipo Gibbs il processo Stirling-Gamma ammette uno schema ad urne nella forma dell'equazione (2.6). In particolare, $(X_n)_{n \geq 1}$ segue il seguente meccanismo generativo:

$$P(X_{n+1} \in \cdot | X_1, \dots, X_n) = \frac{V_{a,b,m}(n+1, H+1)}{V_{a,b,m}(n, H)} P_0(\cdot) + \frac{V_{a,b,m}(n+1, H)}{V_{a,b,m}(n, H)} \sum_{h=1}^H n_h \delta_{X_h^*}(\cdot),$$

per $n \geq 1$.

Gli iperparametri della distribuzione a priori Stirling-Gamma sono interpretabili in termini di numero di gruppo indotto nella partizione latente. Infatti, se $\alpha \sim Sg(a, b, m)$ e $D_{a,b,m} = E\{\sum_{i=0}^{m-1} \alpha^2/(\alpha+i)^2\}$. Il numero di gruppi H_m ottenuti dalle prime m variabili casuali $\theta_1, \dots, \theta_m$ generate dallo schema predittivo sopra descritto ha distribuzione

$$P(K_m = H) = \frac{V_{a,b,m}(m, H)}{V_{a,b,m}(1, 1)} |(s(m, H))|,$$

per $H = 1, \dots, m$, con media e varianza pari a

$$E(K_m) = \frac{a}{b} \quad \text{var}(K_m) = \frac{b+1}{b} \left(\frac{a}{b} - D_{a,b,m} \right).$$

Gli iperparametri a, b e m hanno un importante significato: quando $\alpha \sim Sg(a, b, m)$, le prime m unità statistiche $\{1, \dots, m\}$ provenienti da un processo Stirling-Gamma, identificano in media $\frac{a}{b}$ gruppi, con varianza inversamente proporzionale a b . Per questo motivo, possiamo riferirci a m come l'ipotetica dimensione del campione di riferimento, $\frac{a}{b}$ come locazione e b come precisione. Grazie alla distribuzione del numero di gruppi, si evince una motivazione esplicita del perché gli iperparametri della Stirling-Gamma debbano soddisfare la condizione $1 < \frac{a}{b} < m$: avere $\frac{a}{b} = 1$

comporta che $E(K_m) = 1$, ciò corrisponde a un Processo di Dirichlet dove $\alpha \rightarrow 0$. Al contrario, porre $\frac{a}{b} = m$ porta a $E(K_m) = m$, che vuol dire che ogni osservazione identifica un nuovo cluster. Questo è il caso del Processo di Dirichlet per $\alpha \rightarrow \infty$. Porre $1 < \frac{a}{b} < m$ evita entrambi questi comportamenti degeneri.

Come già evidenziato all'inizio, la distribuzione Stirling-Gamma è coniugata alla legge della partizione del Processo Dirichlet, quando la dimensione del campione di riferimento m è posta uguale al numero di punti n . Supponiamo di osservare una partizione Π_n distribuita secondo il Processo di Dirichlet e sia $\alpha \sim Sg(a, b, m)$, allora:

$$(\alpha \mid \Pi_n = \{S_1, \dots, S_H\}) \sim Sg(a + H, b + 1, n)$$

Tale coniugazione semplifica i calcoli durante il campionamento della distribuzione a posteriori del modello di mistura processo di Dirichlet con precisione casuale. Sotto la distribuzione a priori coniugata Stirling-Gamma l'elicitazione è diretta, per questo motivo è possibile sintonizzare la distribuzione a priori Stirling-Gamma, sfruttando l'informazione disponibile sulla struttura dei gruppi delle n osservazioni, attraverso la scelta di a e b .

2.1.6 SALSO

Concludiamo questa prima parte con la trattazione dell'algoritmo SALSO, introdotto da [Dahl et al. \(2022\)](#), in quanto sarà l'algoritmo che impiegheremo in fase di analisi della partizione ottima.

Andiamo dunque a trattare metodi di clustering per partizioni latenti delle osservazioni, in particolare il nostro obiettivo consiste nel fornire tale classificazione ricorrendo all'utilizzo dei modelli per blocchi stocastici estesi. L'inferenza nei modelli ESBM ricorre ai metodi Monte Carlo Markov Chains (MCMC), che producono un gran numero di partizioni che rappresentano campioni approssimati della distribuzione a posteriori. A causa della grande dimensione dello spazio delle partizioni e al fatto che molte di queste differiscono tra di loro solo in pochi punti, la distribuzione a posteriori solitamente si distribuisce su un gran numero di partizioni. Una partizione $\Pi = \{S_1, \dots, S_p\}$ di interi $\{1, \dots, n\}$ è una collezione di sottoinsiemi mutualmente esclusivi, non vuoti e la cui unione ricostruisce l'intero spazio di partenza. Una partizione può essere alternativamente raffigurata dalle etichette dei cluster, utilizzeremo quindi le due rappresentazioni in maniera interscambiabile. Durante la costruzione di un modello, le partizioni sono spesso utilizzate per organizzare i dati X_1, \dots, X_n in modo tale che essi siano omogenei all'interno di un gruppo.

Gli elementi i e j appartengono allo stesso gruppo se $i \in S$ e $j \in S$ per qualche sottoinsieme $S \in \Pi$.

Come già detto, i metodi MCMC restituiscono un elevato numero di partizioni; descrivere ogni singola partizione non è possibile, sono necessari dunque degli strumenti che ci permettano di sintetizzare l'informazione di nostro interesse. Ricorriamo quindi ad una stima puntuale della distribuzione a posteriori, che deve essere in grado di spiegare il suo comportamento. Ricorrendo alla teoria delle decisioni, la stima puntuale è ottenuta specificando la legge di Bayes che minimizza il valore atteso della distribuzione a posteriori della funzione di perdita scelta, ovvero:

$$\hat{\Pi}^* = \arg \min_{\hat{\Pi}} E(L(\Pi, \hat{\Pi})|D),$$

dove D rappresenta i dati e $L(\Pi, \hat{\Pi})$ è la funzione di perdita. Questo comporta la scelta di una funzione di perdita appropriata per lo spazio dei clustering. Inoltre, è necessario fornire una misura di incertezza attorno alla stima puntuale, questo è possibile costruendo un *credible ball* attorno ad essa.

Studieremo due diverse funzioni di perdita e le loro generalizzazioni per ottenere una stima puntuale e attorno ad essa andremo a caratterizzare l'incertezza. Una scelta semplice sarebbe la perdita $L_{0/1}$, che assume 0 se la stima corrisponde al valore vero, altrimenti vale 1; questa funzione restituisce come stima ottimale la moda a posteriori. Questa funzione di perdita però, non è soddisfacente in quanto non tiene conto delle similarità tra i cluster e inoltre non è detto che sia rappresentativa della centralità della distribuzione. È necessario dunque trovare una funzione di perdita più generale. Secondo Binder, una funzione di perdita deve soddisfare alcuni principi base, come l'invarianza a permutazione degli indici dei punti e delle etichette nei cluster, sia per gli insiemi veri che per quelli stimati. La prima condizione implica che la perdita sia una funzione dei conteggi:

$$n_{i,j} = |S_i \cup \hat{S}_j|,$$

ovvero $n_{i,j}$ indica la cardinalità dell'intersezione tra S_i e \hat{S}_j . Dall'esplorazione di funzioni di perdita che soddisfano queste proprietà, viene formulata la funzione di perdita di Binder, che prende il nome dal suo ideatore.

Alla luce di quanto detto in precedenza, essendoci un'equivalenza tra la rappresentazione in termini di partizione e in termini di etichette dei cluster, introduciamo la seguente notazione: indichiamo con $\mathbf{z} = z_1, \dots, z_n$ il clustering della popolazione e

con $\hat{\mathbf{z}} = \hat{z}_1, \dots, \hat{z}_n$ la sua stima. Possiamo perciò definire la funzione di perdita di Binder come:

$$L_{\text{binder}}(\mathbf{z}, \hat{\mathbf{z}}) = \sum_{i < j} (a \mathbb{1}\{z_i = z_j\} \mathbb{1}\{\hat{z}_i \neq \hat{z}_j\} + b \mathbb{1}\{z_i \neq z_j\} \mathbb{1}\{\hat{z}_i = \hat{z}_j\}),$$

dove $a > 0$ e $b > 0$ rappresentano il costo unitario di ogni errata classificazione, in particolare: a rappresenta l'allocazione di due osservazioni in cluster diversi quando dovrebbero essere nello stesso, mentre b rappresenta l'allocazione nello stesso cluster di due osservazioni quando queste dovrebbero essere separate. [Wade & Ghahramani \(2018\)](#) ne hanno introdotto una versione n-invariante che assume $a = b = 1$:

$$L_{n'\text{binder}}(\Pi, \hat{\Pi}) = \frac{2}{n^2} L_{\text{binder}}(\Pi, \hat{\Pi}) = \sum_{S \in \Pi} \left(\frac{|S|}{n} \right)^2 + \sum_{\hat{S} \in \hat{\Pi}} \left(\frac{|\hat{S}|}{n} \right)^2 - 2 \sum_{S \in \Pi} \sum_{\hat{S} \in \hat{\Pi}} \left(\frac{|S \cup \hat{S}|}{n} \right)^2.$$

Sotto questa funzione di perdita, la partizione ottima \mathbf{z}^* è la partizione che minimizza:

$$\sum_{i < j} (\mathbb{1}(z_i = z_j) - p_{ij})^2,$$

dove $p_{ij} = P(c_i = c_j | X_{1:n})$ è la probabilità a posteriori che due osservazioni i e j siano classificate insieme. Il punto critico di questa funzione di perdita è che tende a sovrastimare il numero di cluster, questo perché preferisce dividerli piuttosto che unirli. Per sopperire a tale problema, [Dahl et al. \(2022\)](#) introduce un'espressione n-invariante della funzione di perdita di Binder utilizzando pesi generici a e b .

$$L_{Gn'\text{binder}}(\Pi, \hat{\Pi}) = a \sum_{S \in \Pi} \left(\frac{|S|}{n} \right)^2 + b \sum_{\hat{S} \in \hat{\Pi}} \left(\frac{|\hat{S}|}{n} \right)^2 - (a + b) \sum_{S \in \Pi} \sum_{\hat{S} \in \hat{\Pi}} \left(\frac{|S \cup \hat{S}|}{n} \right)^2,$$

dove viene infatti mostrato come, ponendo $a > b$ sia possibile controllare il numero dei cluster nella stima.

Un'altra possibilità è l'utilizzo della Variazione d'Informazione (VI) introdotta da Meilă come misura di distanza tra due partizioni e riproposta da [Wade & Ghahramani \(2018\)](#) come funzione di perdita nella stima di una partizione aleatoria. Questa funzione ricorre all'entropia, indicata con $H(\cdot)$, e all'informazione mutua, denotata attraverso $I(\cdot, \cdot)$.

$$\begin{aligned} L_{VI} = (\Pi, \hat{\Pi}) &= H(\Pi) + H(\hat{\Pi}) - 2\mathbb{1}(\Pi, \hat{\Pi}) \\ &= -H(\Pi) - H(\hat{\Pi}) + 2H(\Pi, \hat{\Pi}) \end{aligned}$$

$$= \sum_{S \in \Pi} \frac{|S|}{n} \log_2\left(\frac{|S|}{n}\right) + \sum_{\hat{S} \in \hat{\Pi}} \frac{|\hat{S}|}{n} \log_2\left(\frac{|\hat{S}|}{n}\right) - 2 \sum_{s \in \Pi} \sum_{\hat{S} \in \hat{\Pi}} \frac{|S \cup \hat{S}|}{n} \log_2\left(\frac{|S \cup \hat{S}|}{n}\right).$$

La partizione ottima che suggeriscono sfrutta la disuguaglianza di Jensen ed è la seguente:

$$\hat{\mathbf{z}}^* = \operatorname{argmin}_{\hat{\mathbf{z}}} \sum_{i=1}^n \log_2\left(\sum_{j=1}^n \mathbb{1}(\hat{z}_j = \hat{z}_i)\right) - 2 \sum_{i=1}^n \left(\log_2\left(\sum_{j=1}^n \hat{\pi}_{ij} \mathbb{1}(\hat{z}_j = \hat{z}_i)\right) \right),$$

dove $\hat{\pi}_{ij} = \frac{1}{M} \sum_{m=1}^M \mathbb{1}(z_i^{(m)} = z_j^{(m)})$ e M indica il campione Monte Carlo di riferimento. Questa funzione di perdita in alcuni casi può produrre meno cluster di quanto sia ragionevole, per questo Dahl et al. (2022) propongono una sua versione generalizzata con pesi positivi a e b .

$$\begin{aligned} L_{GVI} &= (\Pi, \hat{\Pi}) = aH(\Pi) + bH(\hat{\Pi}) - (a+b)\mathbb{1}(\Pi, \hat{\Pi}) \\ &= -aH(\Pi) - bH(\hat{\Pi}) + (a+b)H(\Pi, \hat{\Pi}) \\ &= a \sum_{S \in \Pi} \frac{|S|}{n} \log_2\left(\frac{|S|}{n}\right) + b \sum_{\hat{S} \in \hat{\Pi}} \frac{|\hat{S}|}{n} \log_2\left(\frac{|\hat{S}|}{n}\right) - (a+b) \sum_{s \in \Pi} \sum_{\hat{S} \in \hat{\Pi}} \frac{|S \cup \hat{S}|}{n} \log_2\left(\frac{|S \cup \hat{S}|}{n}\right). \end{aligned}$$

In particolare, mostrano come il controllo sul numero dei cluster avvenga nel caso di $a < b$.

Una proprietà desiderabile per una funzione di perdita è che sia metrica. Per soddisfarla, una misura di distanza deve rispettare i seguenti requisiti:

- principio di identità degli indiscernibili (principio ontologico che afferma che due oggetti separati non possono avere tutte le proprietà in comune);
- simmetria;
- disuguaglianza triangolare.

Le due funzioni di perdita proposte con $a = b$ soddisfano tutte queste proprietà, mentre la loro versione generalizzata non soddisfa la proprietà di simmetria, per questo è detta “quasi metrica”. Poichè la funzione di perdita di Binder e la Variazione d’Informazione descritte da Wade & Ghahramani (2018) sono delle metriche nello spazio, è possibile costruire un “credible ball” attorno alla stima. Per caratterizzare l’incertezza nella stima puntuale \mathbf{z}^* , costruiamo un intervallo sferico di livello $1 - \alpha$ con $\alpha \in [0, 1]$ definito come:

$$B_{\epsilon^*}(\mathbf{z}^*) = \{\mathbf{z} : d(\mathbf{z}^*, \mathbf{z}) \leq \epsilon^*\},$$

dove ϵ^* è il più piccolo $\epsilon > 0$ tale che $P(B_\epsilon(\mathbf{c}^*) \mid D) \geq 1 - \alpha$. Il *credible ball* è, dunque, la più piccola sfera attorno a z^* con probabilità a posteriori almeno $1 - \alpha$. Questo riflette l'incertezza a posteriori nella stima puntuale z^* : con probabilità $1 - \alpha$ crediamo che la distanza fra il clustering e z^* sia inferiore a ϵ^* date le osservazioni. Tipicamente i *credible ball* sotto Variazione d'Informazione e funzione di perdita di Binder differiscono.

Concludiamo questa sezione descrivendo l'algoritmo SALSO, introdotto da [Dahl et al. \(2022\)](#), poiché sarà il metodo che impiegheremo per trovare la partizione ottima in fase di analisi, in quanto permette di ammettere un numero maggiore di funzioni di perdita ed è più rapido dell'algoritmo introdotto da [Wade & Ghahramani \(2018\)](#). Questo algoritmo è definito “avid”, si tratta infatti di una procedura che ad ogni step prende piccoli aggiornamenti localmente ottimali lungo la strada per trovare la soluzione finale. Un vantaggio di questa tipologia di algoritmi è che invece di limitarsi alle partizioni visitate nella catena, possono considerare anche partizioni non esplorate dall'algoritmo MCMC; infatti, nella maggior parte degli esempi simulati e reali, il clustering stimato non è tra le partizioni campionate e risulta avere una perdita attesa più bassa di ogni partizione campionata. Entriamo adesso nel dettaglio di questo algoritmo, descrivendo le 4 fasi di cui si compone.

- *Inizializzazione*: L'algoritmo SALSO fornisce due metodi per inizializzare una partizione. Il primo è un'allocazione sequenziale in cui gli oggetti vengono allocati uno alla volta, in un ordine determinato da una permutazione campionata in modo uniforme tra tutte le possibili, in un cluster esistente o in uno nuovo sulla base dell'allocazione che minimizza la stima Monte Carlo della perdita attesa a posteriori, ignorando ogni oggetto ancora da allocare. Nel secondo metodo di inizializzazione, le etichette dei cluster sono ottenute campionando uniformemente le etichette $1, \dots, k_d$, dove k_d indica il numero massimo di cluster desiderato. Il metodo di inizializzazione è scelto casualmente.
- “Sweetening”: In questa fase vengono attuate ulteriori allocazioni di oggetti individuali, in modo casuale, uno alla volta in un ordine casuale. Questa è essenzialmente la stessa idea del metodo di allocazione sequenziale, tranne che, adesso, tutti gli oggetti sono allocati e ogni oggetto, uno alla volta nell'ordine determinato da una permutazione campionata uniformemente tra tutte le possibili, è rimosso dal suo cluster e allocato un'altra volta in un cluster esistente o in uno nuovo secondo la scelta che minimizza la stima Monte Carlo della perdita attesa a posteriori. Questo processo è ripetuto fino a quando non

avviene nessun cambiamento dopo un passaggio completo attraverso tutti gli n oggetti.

- “*Zealous update*”: Cerca di uscire dai minimi locali, questi aggiornamenti possono essere dispendiosi a livello computazionale. Per un ordinamento casuale dei cluster, lo stato corrente è memorizzato e il cluster è distrutto rimuovendo tutti i suoi elementi. Questi oggetti dislocati sono poi sequenzialmente allocati nuovamente, uno alla volta in modo casuale, condizionatamente a quelli già allocati, nello stesso modo dell’allocazione sequenziale nella fase di inizializzazione. Una volta che tutto è stato nuovamente allocato, la stima Monte Carlo della perdita attesa a posteriori della partizione corrente è comparata a quella prima della distruzione del cluster. Se non viene trovato nessun miglioramento in questo aggiornamento, questo viene abbandonato e lo stato ritorna al suo stato precedente.
- *Memorizzazione (“Recording”)*: La stima Monte Carlo della perdita attesa a posteriori è memorizzata per lo stato corrente.

Questo algoritmo è implementato per varie funzioni di perdita, in particolare possiamo utilizzare tutte quelle descritte in precedenza.

La complessità dell’algoritmo SALSO è $O(Mk_d k_H n)$ dove M è il numero di campioni MCMC, k_H è il massimo numero di cluster osservati in essi e n è il numero di elementi.

2.2 Modelli per reti

Passiamo adesso alla descrizione dei modelli impiegati per dati di rete, in particolare andiamo prima a descrivere il caso base dei modelli per blocchi stocastici (SBM) e in seguito illustriamo la loro estensione (ESBM) che, successivamente, sarà impiegata per comprendere la struttura interna de *La Lombardia*.

2.2.1 Modelli per blocchi stocastici

I modelli per blocchi stocastici, trattati da Nowicki & Snijders (2001) sono una classe di modelli impiegata nell’analisi statistica di grafi e reti. Questi possono essere utilizzati per scoprire e comprendere la struttura latente di una rete o per fare clustering.

L'obiettivo dei modelli per blocchi stocastici è quello di fare inferenza contemporaneamente sull'appartenenza ai gruppi e sulla matrice dei blocchi. Di conseguenza sorgono le seguenti sfide statistiche:

- modellazione: come i SBM dovrebbero essere strutturati o estesi per descrivere reti reali e l'eventuale aggiunta di ulteriori informazioni sui nodi o sui legami;
- inferenza: come fare inferenza sull'appartenenza ai gruppi e sulla matrice dei blocchi, una volta calcolata la verosimiglianza;
- selezione e diagnostica: calcolo dei criteri d'informazione per individuare il modello migliore;
- numero di gruppi H : questo non è noto prima della modellazione e dell'inferenza, bisogna decidere se H sia un parametro del modello su cui fare inferenza, o se invece vadano stimati diversi SBM per differenti valori fissi di H così da trovare il suo valore ottimo come problema di selezione del modello.

Andiamo adesso a elencare la notazione necessaria per formulare una versione base dei modelli per blocchi stocastici e ad introdurre il concetto di equivalenza stocastica.

Introduciamo la terminologia base di questi modelli, in parte già trattata durante la descrizione della *rete Infinito*. Consideriamo un grafo non orientato $\mathcal{G} = (W, \mathcal{E})$, dove W è l'insieme dei nodi di dimensione $n := |W|$ e \mathcal{E} indica la lista dei legami di dimensione $M := |\mathcal{E}|$. Consideriamo la presenza o l'assenza di un legame tra una coppia di nodi u e v attraverso l'utilizzo della matrice di adiacenza $n \times n$, indicata con \mathbf{Y} tale che:

$$Y_{uv} = \begin{cases} 1 & \text{se } u \text{ e } v \text{ hanno un legame tra di loro,} \\ 0 & \text{altrimenti,} \end{cases}$$

inoltre per costruzione \mathbf{Y} è simmetrica lungo la diagonale principale.

In questi modelli ogni nodo appartiene a un gruppo $H(< n)$. Poiché i gruppi non sono noti prima della modellazione, per il nodo $v = 1, 2, \dots, n$ è definito il vettore \mathbf{z}_v , vettore di lunghezza H dove tutti gli elementi sono 0, eccetto uno che assume valore 1, questo rappresenta il gruppo d'appartenenza del nodo v . Definiamo inoltre la matrice $n \times H$ $\mathbf{Z} := (\mathbf{z}_1, \dots, \mathbf{z}_n)^T$, tale che z_{vi} è l' i -esimo elemento di \mathbf{z}_v . La dimensione dei gruppi può essere derivata da \mathbf{Z} e possiamo indicarla con $\mathbf{N} = (N_1, \dots, N_H)^T$, dove N_h somma gli elementi diversi da 0 in \mathbf{z}_h . Infine è possibile derivare da \mathbf{Z} e \mathbf{Y} la

matrice dei legami \mathbf{E} , dove $E_{h,k}$ rappresenta il numero di legami tra i gruppi h e k . È possibile descrivere la generazione dei legami del grafo \mathcal{G} secondo i gruppi a cui appartengono i nodi attraverso la matrice dei blocchi Θ , $H \times H$, dove $\theta_{hk} \in [0, 1]$ rappresenta la probabilità dell'occorrenza di un legame tra un nodo del gruppo h e uno del gruppo k , per $1 \leq h \leq k \leq H$, la matrice Θ è simmetrica.

L'idea della matrice a blocchi Θ indica che le coppie di nodi sono condizionatamente indipendenti, data l'appartenenza ai gruppi \mathbf{Z} . In altre parole Y_{vu} segue una distribuzione Bernoulli: $Y_{vu} | z_v, z_u \sim \text{Bernoulli}(\theta_{z_v, z_u})$ e, sempre condizionatamente all'appartenenza ai gruppi z_v, z_u , Y_{vu} è indipendente da Y_{pq} per $(v, u) \neq (p, q)$. Questo implica che il numero di legami tra due gruppi qualsiasi h e k segue una distribuzione Binomiale: (numero totale di legami tra due gruppi h e k) \sim Binomiale (Θ_{hk} · numero di coppie disponibili).

L'assunzione che la probabilità di un legame tra una coppia di nodi dipenda solamente dalla loro appartenenza ai gruppi e da Θ è basata sul concetto di equivalenza stocastica. L'equivalenza stocastica indica che per i nodi v e u appartenenti allo stesso gruppo, v ha la stessa probabilità di connettersi con un nodo p di u , dove tale probabilità dipende dall'appartenenza ai gruppi di p . Il concetto di equivalenza stocastica in sè non richiede che i nodi contenuti nello stesso gruppo siano più connessi tra di loro, rispetto ai nodi degli altri gruppi. In sostanza gli elementi lungo la diagonale principale di Θ non sono necessariamente più alti degli elementi fuori dalla diagonale, anche se tale fenomeno è spesso l'obiettivo quando si vuole individuare una comunità.

Infine formuliamo il modello sottostante i modelli per blocchi stocastici. Tali modelli partizionano i nodi in H gruppi mutualmente esclusivi ed esaustivi, dove i nodi appartenenti al medesimo gruppo hanno in comune lo stesso schema di connessioni. In particolare i modelli per blocchi stocastici estesi assumono che le entrate della sotto-diagonale Y_{vu} della matrice di adiacenza \mathbf{Y} siano condizionatamente variabili casuali Bernoulli indipendenti con probabilità associate $\theta_{\mathbf{z}_v}, \theta_{\mathbf{z}_u} \in (0, 1)$, che dipendono solo dall'appartenenza ai gruppi \mathbf{z}_v e \mathbf{z}_u dei nodi coinvolti v ed u .

D'ora in avanti indicheremo con $\mathbf{z} = (z_1, \dots, z_W)^T \in \{1, \dots, H\}$ il vettore d'appartenenza dei nodi, associato ad una generica partizione dei nodi $\{S_1, \dots, S_h\}$, tale che $z_v = h$ se e solo se $v \in S_h$. Inoltre ricordiamo che Θ indica la matrice simmetrica $H \times H$, il cui generico elemento $\theta_{hk} \in (0, 1)$ misura la probabilità che ci sia un legame tra un nodo del gruppo h e un nodo del gruppo k . Allora la

verosimiglianza per \mathbf{Y} è:

$$P(\mathbf{Y} \mid \mathbf{z}, \Theta) = \prod_{h=1}^H \prod_{k=1}^h \theta_{hk}^{m_{hk}} (1 - \theta_{hk})^{\bar{m}_{hk}},$$

dove m_{hk} e \bar{m}_{hk} indicano rispettivamente il numero di legami e di non legami tra i nodi del gruppo h e quelli appartenenti al gruppo k .

I modelli per blocchi stocastici classici assumono una distribuzione a priori indipendente $Beta(a, b)$ per le probabilità dei blocchi Θ . Perciò la densità congiunta per gli elementi della diagonale e della sotto-diagonale di Θ è:

$$p(\Theta) = \prod_{h=1}^H \prod_{k=1}^h \frac{\theta_{hk}^{a-1} (1 - \theta_{hk})^{b-1}}{B(a, b)},$$

dove $B(\cdot, \cdot)$ è la funzione Beta. Nonostante quantificare l'incertezza sia importante, l'obiettivo generale nei modelli per blocchi stocastici è quello di stimare la partizione dei nodi. Coerentemente con tale scopo, Θ viene comunemente trattato come un parametro di disturbo e perciò viene marginalizzato dalla $p(\mathbf{Y}|\mathbf{z}, \Theta)$, attraverso la coniugatezza tra distribuzione beta e binomiale, ottenendo:

$$p(\mathbf{Y}|\mathbf{z}) = \prod_{h=1}^H \prod_{k=1}^h \frac{B(a + m_{hk}, b + \bar{m}_{hk})}{B(a, b)}. \quad (2.9)$$

Tale marginalizzazione è utile anche per calcolare l'inferenza. La verosimiglianza (2.9) è comune a diverse estensioni dei modelli per blocchi stocastici, i quali poi differiscono nella scelta del meccanismo probabilistico sottostante \mathbf{z} .

Sia $\bar{H} \geq H$ il numero totale di gruppi possibili nell'intera popolazione dei nodi e indichiamo con $\bar{\mathbf{z}} = (\bar{z}_1, \dots, \bar{z}_W)^T \in \{1, \dots, H\}^W$ gli indicatori che definiscono i gruppi della popolazione per i W nodi osservati. Un'opzione naturale per caratterizzare il processo generativo per la partizione consiste nel considerare una distribuzione a priori Dirichlet-Multinomiale per $\bar{\mathbf{z}}$, ottenuta attraverso la marginalizzazione del vettore delle probabilità dei gruppi: $\pi = (\pi_1, \dots, \pi_{\bar{H}}) \sim Dirichlet(\beta)$ da una verosimiglianza multinomiale per $\bar{\mathbf{z}}$, in cui $p(\bar{z}_v = h \mid \pi) = \pi_h$, per $v = 1, \dots, W$. Se \bar{H} è fisso e finito si fa riferimento ai modelli per blocchi stocastici bayesiani originali, ma come già detto il numero di gruppi nelle reti criminali non è noto e deve essere stimato dai dati. Una possibile soluzione consiste nel mettere una prior su \bar{H} , questo porta alla versione mistura di misture finite (MFM) dei SBM. Un'altra opzione consiste nell'adottare come meccanismo di partizione il Processo di Dirichlet, questo corrisponde al modello

relazionale infinito. Un tale modello di mistura infinito differisce dalla MFM per $\bar{H} = \infty$, ciò vuol dire che un'infinità di nodi dà origine ad un'infinità di gruppi. Sottolineiamo che il numero totale di possibili gruppi \bar{H} non deve essere confuso con il numero di gruppi occupati H . Quest'ultimo è definito come il numero di etichette distinte in $\bar{\mathbf{z}}$ e il suo limite superiore è $\min\{W, \bar{H}\}$. Possiamo notare come tutte le soluzioni sovra elencate siano esempi specifici delle distribuzioni a priori di tipo Gibbs, questo motiva l'utilizzo di tale classe unificata nei modelli per blocchi stocastici estesi.

Osserviamo che $\bar{\mathbf{z}}$ identifica le etichette dei gruppi, pertanto un vettore $\bar{\mathbf{z}}$ e la sua permutazione delle etichette sono viste come oggetti distinti, anche se identificano la stessa partizione. D'ora in avanti ci riferiremo al vettore \mathbf{z} , definito precedentemente, per indicare tutte le permutazioni delle etichette di $\bar{\mathbf{z}}$ che portano alla stessa partizione. Per semplicità assumiamo che $z_v \in \{1, \dots, H\}$, che corrisponde ad evitare gruppi vuoti. Questo non modifica la verosimiglianza (2.9), la quale è invariante a permutazioni delle etichette, ovvero $p(\mathbf{Y}|\mathbf{z}) = p(\mathbf{Y} | \bar{\mathbf{z}})$.

2.2.2 Modelli per blocchi stocastici estesi

Andiamo adesso ad estendere la classe di modelli appena vista, così da ottenere i modelli per blocchi stocastici estesi

Come abbiamo mostrato durante la trattazione dei modelli per blocchi stocastici, diverse distribuzioni a priori sono state considerate per \mathbf{z} , tra cui la Dirichlet-Multinomiale, il Processo di Dirichlet e le misture di misture finite di Dirichlet. Questi sono tutti esempi di distribuzioni a priori di tipo Gibbs, illustrate in precedenza. I modelli per blocchi stocastici estesi costituiscono una struttura unificatrice e sono caratterizzati dalla scelta di distribuzioni a priori di tipo Gibbs. Tale formulazione include i modelli per blocchi stocastici come caso speciale e offre nuove alternative esplorando tutta la classe di distribuzioni a priori di tipo Gibbs e anche grazie alla sua diretta relazione con i modelli per prodotti di partizioni (PPM). Questa connessione con i PPM verrà sfruttata per controllare la distribuzione a priori attraverso gli attributi dei nodi, i quali sono possibilmente soggetti ad errore.

Andiamo adesso a caratterizzare i modelli per blocchi stocastici estesi. Gli indicatori dei gruppi non etichettati \mathbf{z} seguono una distribuzione a priori di tipo Gibbs, definita nella (2.5). Inoltre abbiamo visto come la classe delle partizioni casuali indotte dalle distribuzioni a priori di tipo Gibbs coincide con i PPM scambiabili che, in questo contesto, sono i modelli di probabilità per le partizioni casuali, per

questo $p(\mathbf{z})$ è proporzionale alla (2.7). Mostreremo come tale connessione sia utile per incorporare gli attributi specifici dei nodi nei modelli per blocchi stocastici estesi.

Le distribuzioni a priori di tipo Gibbs rappresentano una classe ampia, ma trattabile come abbiamo visto, la cui distribuzione predittiva implica che gli indicatori d'appartenenza ai gruppi \mathbf{z} possono essere ottenuti in modo sequenziale ed interpretabile, secondo la (2.8). Tale meccanismo generale è concettualmente interessante nella nostra applicazione a reti criminali dal momento in cui tiene conto, verosimilmente, della dimensione del gruppo n_h , della dimensione della rete W e dalla complessità H nel processo di formazione della struttura in moduli sottostante ad uno schema di modellazione singolo.

Se sono disponibili gli attributi dei nodi $\mathbf{x}_v = (x_{v1}, \dots, x_{vd})^T$ per ogni $v = 1, \dots, W$, questa informazione esogena potrebbe supportare l'inferenza sulle strutture dei blocchi, sia in termini di stima puntuale che in riduzione dell'incertezza a posteriori. Come abbiamo detto in precedenza, questo è particolarmente rilevante nelle applicazioni alle reti criminali, dove la struttura specifica dei blocchi può essere offuscata di proposito da strategie volte a nascondere la vera composizione della rete e, perciò, l'inclusione di attributi informativi potrebbe aiutare a svelare moduli nascosti. Questa soluzione dovrebbe anche tenere conto del fatto che gli attributi dei nodi collezionati durante le investigazioni potrebbero essere soggetti ad errore. Un'opzione per affrontare tale obiettivo all'interno dei modelli per blocchi stocastici estesi consiste nel fare affidamento sulla struttura dei modelli per prodotti di partizioni (PPM) delle priori di tipo Gibbs. La soluzione è basata sulla sostituzione di (??) con

$$p(\mathbf{z} | \mathbf{X}) \propto V_{W,H} \prod_{h=1}^H p(\mathbf{X}_h)(1 - \sigma)_{n_h - 1}, \quad (2.10)$$

dove $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$, mentre $\mathbf{X}_h = \{\mathbf{x}_v : z_v = h\}$ sono gli attributi dei nodi del gruppo h . Nella formulazione (2.10) $p(\mathbf{X}_h)$ controlla il contributo di \mathbf{X} alla coesione tra gruppi, favorendo gruppi che sono omogenei rispetto ai valori degli attributi, includendo allo stesso tempo dell'incertezza su \mathbf{X} . Motivata dall'applicazione alla *rete Infinito*, consideriamo il caso in cui ogni attributo dei nodi $\mathbf{x}_v = x_v \in \{1, \dots, C\}$ è una singola variabile categoriale che indica una combinazione appropriata tra la *locale* d'appartenenza e il ruolo nell'organizzazione criminale. Questa composizione è comune in criminologia, dove gli attributi dei nodi spesso provengono da partizioni esogene definite dalle forze dell'ordine come risultato del processo di investigazione. In questo contesto categoriale ricorrendo alla struttura dei PPM, è raccomandato utilizzare la

funzione di coesione Dirichlet-Multinomiale, senza il coefficiente multinomiale:

$$p(\mathbf{X}_h) \propto \frac{1}{\Gamma(n_h + \alpha_0)} \prod_{c=1}^C \Gamma(n_{hc} + \alpha_c), \quad (2.11)$$

dove n_{hc} è il numero di nodi nel gruppo h con valore dell'attributo c , mentre $\alpha_0 = \sum_{c=1}^C \alpha_c$, con $\alpha_c > 0$ per $c = 1, \dots, C$. Includere questa funzione di coesione nella (2.10) porta al seguente schema ad urna:

$$p(z_{W+1} = h \mid \mathbf{z}, x_{W+1}, \mathbf{z}) \propto \begin{cases} \frac{n_{hx_{W+1}} + \alpha_{x_{W+1}}}{n_h + \alpha_0} V_{W+1,H}(n_h - \sigma) & \text{per } h = 1, \dots, H, \\ \frac{\alpha_{x_{W+1}}}{\alpha_0} V_{W+1,H+1} & \text{per } h = H+1, \end{cases} \quad (2.12)$$

dove $n_{hx_{W+1}}$ è il numero di nodi nel gruppo h che hanno lo stesso valore della covariata $c = x_{W+1}$ come il nodo $W+1$, n_h è il numero totale di nodi nel gruppo h , mentre $\alpha_{x_{W+1}}$ è il parametro associato alla categoria $c = x_{W+1}$ del nodo $W+1$. Come viene chiarito nella (2.12), l'introduzione della $p(\mathbf{X}_h)$, induce un'omofilia probabilistica favorendo l'attribuzione di un nuovo nodo a quei gruppi che hanno una frazione più alta di nodi esistenti con lo stesso valore dell'attributo. Oltre ad includere strutture di omofilia realistiche, la rappresentazione sovrastante tiene effettivamente conto del possibile rumore degli attributi. Infatti l'espressione per $p(\mathbf{X}_h)$ nella (2.11) coincide con la verosimiglianza marginale per gli attributi dei nodi nel gruppo h , sotto l'assunzione che il modello sottostante tutte queste quantità sia definito da una multinomiale con probabilità delle classi specifiche per gruppo: $\pi_h = (\pi_{1h}, \dots, \pi_{Ch})^T$, a cui è assegnata una distribuzione a priori Dirichlet con parametri $\alpha = (\alpha_1, \dots, \alpha_C)^T$. Sotto questa interpretazione, le prior di tipo Gibbs supervisionate nell'equazione (2.10) possono essere riscritte come

$$p(\mathbf{z} \mid \mathbf{X}) \propto [V_{n,H} \prod_{h=1}^H (1 - \sigma)_{n_h-1}] \prod_{h=1}^H p(\mathbf{X}_h) \propto p(\mathbf{z}) p(\mathbf{X} \mid \mathbf{z}),$$

dove $p(\mathbf{z})$ è la prior di tipo Gibbs non supervisionata, mentre $p(\mathbf{X} \mid \mathbf{z})$ è la verosimiglianza indotta dal modello Dirichlet-Multinomiale per gli attributi dei nodi osservati. Perciò imparare le strutture dei blocchi in \mathbf{Y} sotto le prior supervisionate di tipo Gibbs può essere interpretata come una procedura bayesiana in due stadi. In primo luogo la prior non supervisionata \mathbf{z} è aggiornata con la verosimiglianza degli attributi in \mathbf{X} e poi la posterior, ottenuta in questo primo step, entra come una nuova prior nel secondo stadio, aggiornando l'informazione proveniente dalla rete

osservata \mathbf{Y} . Sotto l'assunzione dell'indipendenza condizionata tra \mathbf{Y} e \mathbf{X} dato \mathbf{z} , questa procedura a due stadi porta all'effettiva prior per \mathbf{z} , dal momento in cui:

$$p(\mathbf{z} \mid \mathbf{Y}, \mathbf{X}) \propto [p(\mathbf{z})p(\mathbf{X} \mid \mathbf{z})]p(\mathbf{Y} \mid \mathbf{z}) \propto p(\mathbf{z} \mid \mathbf{X})p(\mathbf{Y} \mid \mathbf{z}).$$

Andiamo adesso a derivare un campionamento Gibbs collassato, il quale può essere impiegato per l'intera classe dei modelli per blocchi stocastici estesi. La disponibilità degli schemi ad un urne nella (??) e nella (2.12) per l'intera classe di prior di tipo Gibbs permette di derivare un campionamento di tipo Gibbs che può essere impiegato per qualsiasi ESBM, questo viene descritto nell'Algoritmo 1. Ad ogni iterazione, questa procedura campiona il gruppo da assegnare per ogni nodo $v = 1, \dots, W$ dalla sua distribuzione full conditional, data la matrice di adiacenza \mathbf{Y} e il vettore \mathbf{z}_{-v} , il quale indica il gruppo assegnato a tutti gli altri nodi, escluso il nodo v . Applicando direttamente la legge di Bayes, queste probabilità delle full conditionals sono:

$$p(z_v = h \mid \mathbf{Y}, \mathbf{X}, \mathbf{z}_{-v}) \propto p(z_v = h \mid \mathbf{X}, \mathbf{z}_{-v}) = \frac{p(\mathbf{Y} \mid z_v = h, \mathbf{z}_{-v})}{p(\mathbf{Y}_{-v} \mid \mathbf{z}_{-v})}, \quad (2.13)$$

dove \mathbf{Y}_{-v} è la matrice di adiacenza $(W - 1) \times (W - 1)$ senza la riga e la colonna che si riferiscono al nodo v . L'ultimo termine in (2.13) può essere semplificato come:

$$\frac{p(\mathbf{Y} \mid z_v = h, \mathbf{z}_{-v})}{p(\mathbf{Y}_{-v} \mid \mathbf{z}_{-v})} = \prod_{k=1}^H \frac{B(a + m_{\bar{h}\bar{k}} + r_{vk}, b + \bar{m}_{\bar{h}\bar{k}} + \bar{r}_{vk})}{B(a + m_{\bar{h}\bar{k}}, b + \bar{m}_{\bar{h}\bar{k}})} \quad (2.14)$$

dove $m_{\bar{h}\bar{k}}$ e $\bar{m}_{\bar{h}\bar{k}}$ indicano rispettivamente il numero di legami e di non legami tra i nodi dei gruppi h e k , senza conteggiare il nodo v ; mentre r_{vk} e \bar{r}_{vk} definiscono rispettivamente il numero di legami e di non legami tra il nodo v e i nodi appartenenti al gruppo k . Il termine a priori $p(z_v = h \mid \mathbf{X}, \mathbf{z}_{-v})$ nella (2.13) è direttamente disponibile dalla (??) e dalla (2.12) a seconda che gli attributi dei nodi siano rispettivamente esclusi o inclusi. In particolare le priori di tipo Gibbs non supervisionate portano a:

$$p(z_v = h \mid \mathbf{X}, \mathbf{z}_{-v}) = p(z_v = h \mid \mathbf{z}_{-v}) \propto \begin{cases} V_{W,H^-}(n_{\bar{h}} - \sigma) & \text{per } h \leq H^-, \\ V_{W,H^+} & \text{per } h = H^+ + 1, \end{cases} \quad (2.15)$$

dove $n_{\bar{h}}$ e H^- indicano rispettivamente la cardinalità del gruppo h e il numero totale di gruppi occupati, dopo aver rimosso il nodo v ; mentre l'estensione al caso

supervisionato porta ad ottenere:

$$p(z_v = h \mid \mathbf{X}, \mathbf{z}_{-v}) \propto \begin{cases} \frac{n_{h\bar{x}_v} + \alpha_{x_v}}{n_h + \alpha_0} V_{W,H}(n_h - \sigma) & \text{per } h \leq H^-, \\ \frac{\alpha_{x_v}}{\alpha_0} V_{W,H+1} & \text{per } h = H^- + 1, \end{cases} \quad (2.16)$$

dove $n_{h\bar{x}_v}$ rappresenta il numero di nodi contenuti nel gruppo h che hanno come valore della covariata $c = x_v$, senza contare il nodo v ; mentre α_{x_v} coincide con il parametro per la categoria $c = x_v$ del nodo v . Sotto le distribuzioni a priori presentate nella sezione riguardante la classe di tipo Gibbs, entrambe la (2.15) e la (2.16) ammettono espressioni semplici.

Algorithm 1: Campionamento Gibbs per ESBM

Ad ogni iterazione aggiornare i gruppi assegnati z_1, \dots, z_W come segue

Per $v = 1, \dots, W$

1. Rimuovere il nodo v dalla rete
 2. Se il gruppo contenente il nodo v diventa vuoto, scartarlo ed etichettare nuovamente gli indicatori dei gruppi (così che i gruppi $1, \dots, H^-$ siano non vuoti)
 3. Campionare z_v dalla variabile casuale categorica con probabilità come nella (2.13) per $h = 1, \dots, H^- + 1$, dove $p(\mathbf{Y} \mid z_v = h, \mathbf{z}_{-v})/p(\mathbf{Y}_{-v} \mid \mathbf{z}_{-v})$ è definita nella (2.14), mentre $p(z_v = h \mid \mathbf{X}, \mathbf{z}_{-v})$ coincide con la (2.15) o la (2.16) a seconda che gli attributi dei nodi siano rispettivamente esclusi o inclusi.
-

Nonostante l'Algoritmo 1 utilizzi la verosimiglianza (2.9) con le probabilità dei gruppi θ_{hk} integrate via, è possibile ottenere facilmente una stima plug-in per ciascun θ_{hk} . In particolare, dal momento in cui $(\theta_{hk} \mid \mathbf{Y}, \mathbf{z}) \sim Beta(a + m_{hk}, b + \bar{m}_{hk})$ una stima puntuale ragionevole per θ_{hk} è:

$$\hat{\theta}_{hk} = E(\theta_{hk} \mid \mathbf{Y}, \mathbf{z} = \hat{\mathbf{z}}) = \frac{a + \hat{m}_{hk}}{a + \hat{m}_{hk} + b + \hat{m}_{hk}},$$

per ogni $h = 1, \dots, \hat{H}$ e $k = 1, \dots, h$, dove \hat{m}_{hk} e \hat{m}_{hk} indicano rispettivamente il numero di legami e di non legami tra i nodi nei gruppi h e k , calcolato dalla stima di $\hat{\mathbf{z}}$.

La selezione di un modello tra diversi candidati è importante nel nostro contesto, perché bisogna scegliere tra le diverse prior di tipo Gibbs prima descritte. Una possibilità per raggiungere tale obiettivo consiste nel ricorrere al fattore di Bayes. Questa strategia, però richiede il calcolo della verosimiglianza marginale:

$p(\mathbf{Y} \mid \mathcal{M}) = \sum_{\mathbf{z}} p(\mathbf{Y} \mid \mathbf{z})p(\mathbf{z} \mid \mathcal{M})$ per un generico modello \mathcal{M} , la quale non è disponibile analiticamente sotto le distribuzioni a priori di tipo Gibbs che abbiamo menzionato. Nonostante alcune semplici strategie, come la stima della media armonica, possano essere impiegate per calcolare la $p(\mathbf{Y} \mid \mathcal{M})$ nei modelli per blocchi stocastici, queste soluzioni in contesti generali potrebbero dover far fronte ad instabilità e lenta convergenza. Per risolvere questi punti deboli e fornire una strategia di selezione dei modelli, ricorriamo al criterio d'informazione WAIC. Rispetto ad altri criteri d'informazione comunemente impiegati nel contesto dei modelli per blocchi stocastici e loro estensioni, il WAIC porta a vantaggi teorici e pratici, inoltre ha una connessione diretta con la validazione incrociata *leave-one-out* bayesiana. Oltretutto, il calcolo del WAIC richiede solo campioni a posteriori della log-verosimiglianza dei legami:

$$\log p(y_{uv} \mid \mathbf{z}, \Theta) = y_{vu} \log \theta_{z_v, z_u} + (1 - y_{vu}), \quad v = 2, \dots, W, \quad u = 1, \dots, v - 1.$$

Queste quantità possono essere facilmente ottenute combinando i campioni a posteriori per \mathbf{z} , ottenuti dall'Algoritmo 1, con quelli delle probabilità a blocchi in Θ , quest'ultimi possono essere prontamente simulati dalla distribuzione coniugata full conditional $(\theta_{hk} \mid \mathbf{Y}, \mathbf{z}) \sim Beta(a + m_{hk}, b + \bar{m}_{hk})$ per $h = 1, \dots, H$ e $k = 1, \dots, h$, attraverso un algoritmo separato che può essere eseguito in parallelo sui blocchi e sui campioni.

Capitolo 3

Valutazione dei ESBM attraverso dati simulati

In questo capitolo andremo ad applicare i modelli per blocchi stocastici estesi, utilizzando le distribuzioni a priori di tipo Gibbs per la $p(\mathbf{z})$, descritte nel precedente capitolo: Dirichlet-Multinomiale, Processo di Dirichlet, Pitman-Yor, Gneden e Stirling-Gamma, su dati simulati per valutarne la performance.

In primo luogo andremo a vedere come questi modelli si comportano al variare del numero di gruppi e delle probabilità contenute nella matrice a blocchi, sottostanti la creazione della rete, inoltre confronteremo i risultati ottenuti con ciascuna distribuzione a priori per diversi valori degli iperparametri. In secondo luogo andremo a simulare una rete che riproduce le caratteristiche della *rete Infinito*, e vedremo, prima senza l'utilizzo di attributi specifici dei nodi e in seguito considerando anche questi, i risultati prodotti dai modelli per blocchi stocastici estesi, così da valutare l'efficacia di questi modelli in un contesto che simula la realtà e vedere fino a che punto le distribuzioni a priori di tipo Gibbs supervisionate riescono a sfruttare tale informazione.

3.1 Applicazione su reti simulate semplici

In questa sezione andremo a simulare quattro matrici di adiacenza differenti, su di esse andremo ad applicare i modelli per blocchi stocastici estesi, per poi analizzare la distribuzione a posteriori del numero di gruppi, confrontare la partizione ottima ottenuta tramite l'algoritmo SALSO con la vera partizione sottostante i dati e guardare quale distribuzione a priori produce il modello migliore per il criterio d'informazione WAIC.

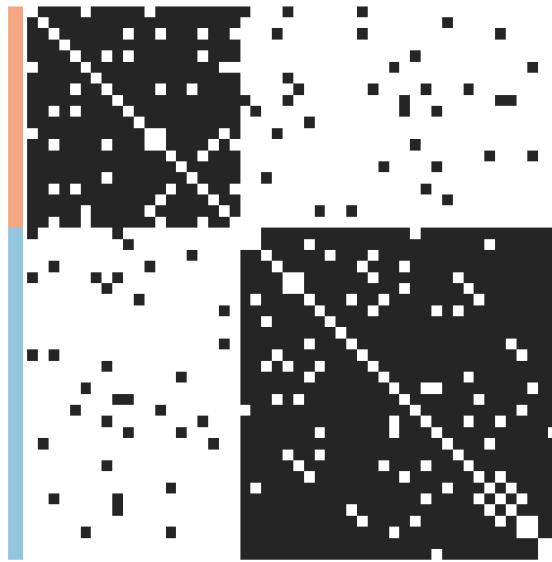


Figura 3.1: Matrice di adiacenza per la rete composta da due gruppi

La prima rete simulata è costituita da $n = 50$ nodi e $H_0 = 2$ gruppi, i primi 20 appartenenti ad un gruppo ed i restanti 30 ad un altro. La matrice a blocchi Θ è caratterizzata da una probabilità di 0.9 che due nodi appartenenti allo stesso gruppo abbiano un legame e di 0.05 che due nodi facenti parte di due gruppi differenti abbiano un legame.

$$\Theta = \begin{bmatrix} 0.9 & 0.05 \\ 0.05 & 0.9 \end{bmatrix}$$

La matrice di adiacenza \mathbf{Y} simulata da questi dati è rappresentata in Figura 3.1, nella quale i colori laterali corrispondono alla vera partizione \mathbf{z}_0 , le celle nere si riferiscono ai legami, mentre quelle bianche ai non legami. Notiamo, attraverso questa raffigurazione, che impiegando la matrice a blocchi Θ , sopra descritta, i due gruppi sono ben distinti tra di loro, questo dovrebbe rendere più semplice l'individuazione della partizione ottima corretta. Utilizziamo due diverse combinazioni degli iperparametri per ciascuna distribuzione a priori di tipo Gibbs elencata in precedenza, la prima che restituisce come valore atteso del numero di gruppi quello effettivo, 2, la seconda che fornisce un numero atteso di gruppi maggiore di 2, per vedere come cambiano i risultati. Iniziamo con l'analisi del primo caso, dove per la distribuzione a priori Dirichlet-Multinomiale impostiamo $\bar{H} = 8$ e $\sigma = -0.0375$, per il Processo di Dirichlet $\alpha = 0.25$, per la distribuzione a priori Pitman-Yor $\sigma = 0.15$ e $\alpha = 0.03$ e per la Gnedin impostiamo $\gamma = 0.8$. Per quanto riguarda la distribuzione a priori Stirling-Gamma, d'ora in avanti ci riferiremo ai suoi parametri a e b rispettivamente con α e β , per non creare confusione, in quanto le lettere a e b in questo contesto le utilizziamo per

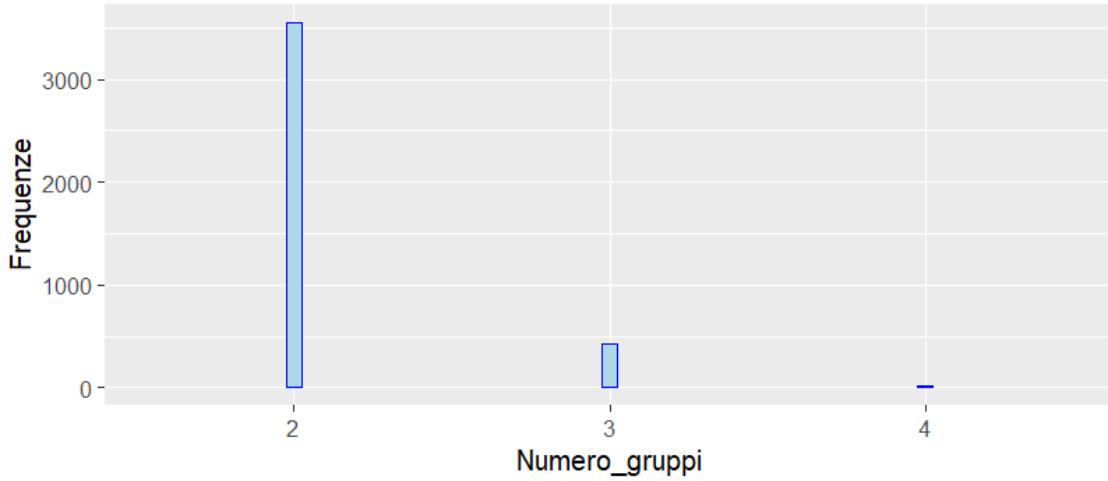


Figura 3.2: Distribuzione a posteriori del numero di gruppi, utilizzando la distribuzione a priori Dirichlet-Multinomiale di parametri $\bar{H} = 8$ e $\sigma = -0.0375$ sulla rete costituita da due gruppi.

indicare i parametri della distribuzione Beta, impostiamo dunque $\alpha = 4$ e $\beta = 2$. Implementiamo i modelli per blocchi stocastici estesi considerando un contesto standard uniforme, impostiamo quindi $a = b = 1$ per le distribuzioni a priori sulle probabilità dei blocchi. Eseguiamo l’Algoritmo 1 e otteniamo 4000 campioni per \mathbf{z} , dopo un *burn-in* di 1000. Per avere un buon *mixing* iniziamo assegnando ogni nodo ad un gruppo differente. Vediamo la distribuzione a posteriori del numero di gruppi per ciascuna di queste distribuzioni a priori attraverso i grafici rappresentati in Figura 3.2, Figura 3.3, Figura 3.4, Figura 3.5 e Figura 3.6. Possiamo notare come per tutte le distribuzioni a priori utilizzate la distribuzione a posteriori del numero di gruppi si concentri sul valore 2, vero numero di gruppi e valore atteso per combinazione dei parametri impiegata. La distribuzione più variabile risulta essere quella ottenute grazie alla distribuzione a priori Stirling-Gamma, dove il numero di gruppi rilevato va da 2 a 5, mentre per tutte le altre distribuzioni va da 2 a 4. Per trovare la partizione ottima tra i 4000 campioni per \mathbf{z} impieghiamo l’algoritmo SALSO con funzione di perdita Binder di parametro $a = 1$. In questo caso la partizione ottima \mathbf{z}^* coincide con la partizione effettiva, per tutte le distribuzioni a priori utilizzate. Abbiamo verificato dunque che in un contesto base, dove la rete in analisi è molto semplice, in quanto costituita da due soli gruppi ben distinti tra di loro, ognuno di questi composto da un numero congruo di nodi e impostando gli iperparametri delle *prior* in modo da ottenere un valore atteso del numero di gruppi che corrisponde a quello effettivo, il risultato ottenuto ricostruisce l’effettiva composizione della rete. Per confrontare le performance delle varie distribuzioni a priori impiegate, le quali hanno

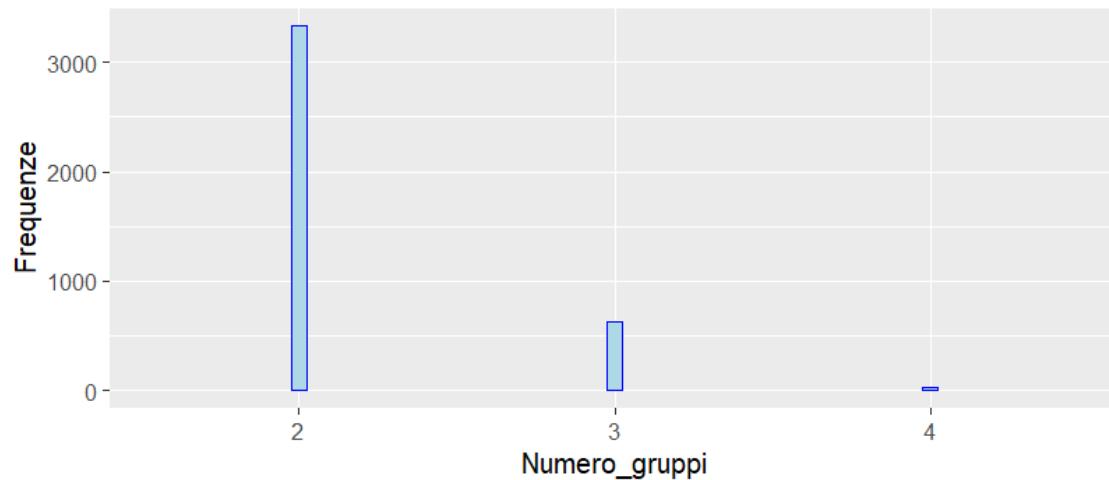


Figura 3.3: Distribuzione a posteriori del numero di gruppi, utilizzando la distribuzione a priori Processo di Dirichlet di parametro $\alpha = 0.25$ sulla rete costituita da due gruppi.

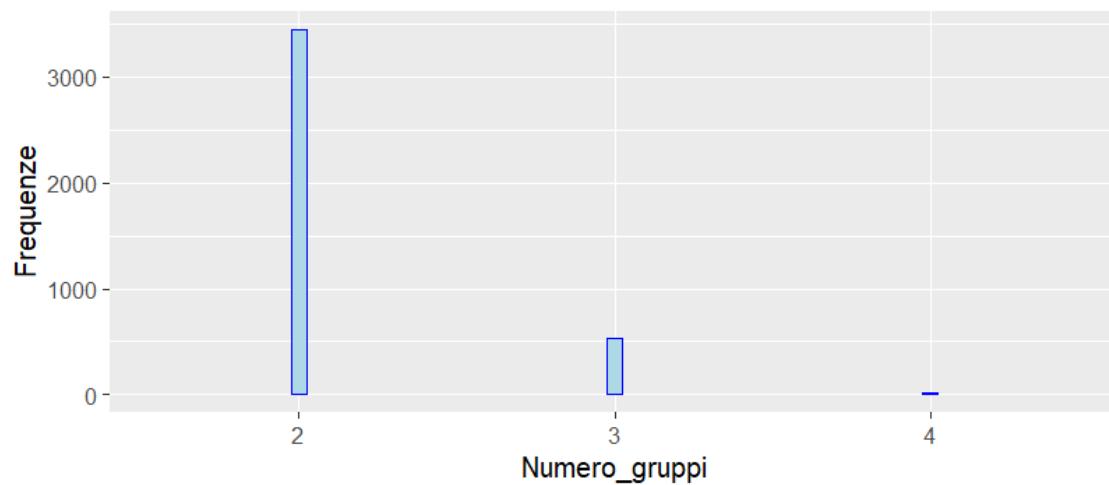


Figura 3.4: Distribuzione a posteriori del numero di gruppi, utilizzando la distribuzione a priori Pitman-Yor di parametri $\sigma = 0.15$ e $\alpha = 0.03$ sulla rete costituita da due gruppi.

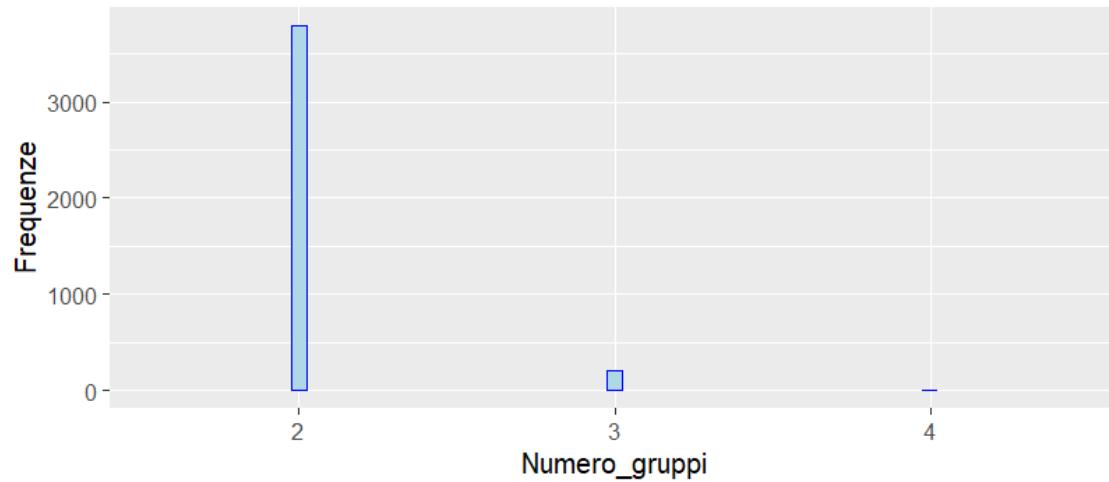


Figura 3.5: Distribuzione a posteriori del numero di gruppi, utilizzando la distribuzione a priori di Gnedin di parametro $\gamma = 0.8$ sulla rete costituita da due gruppi.

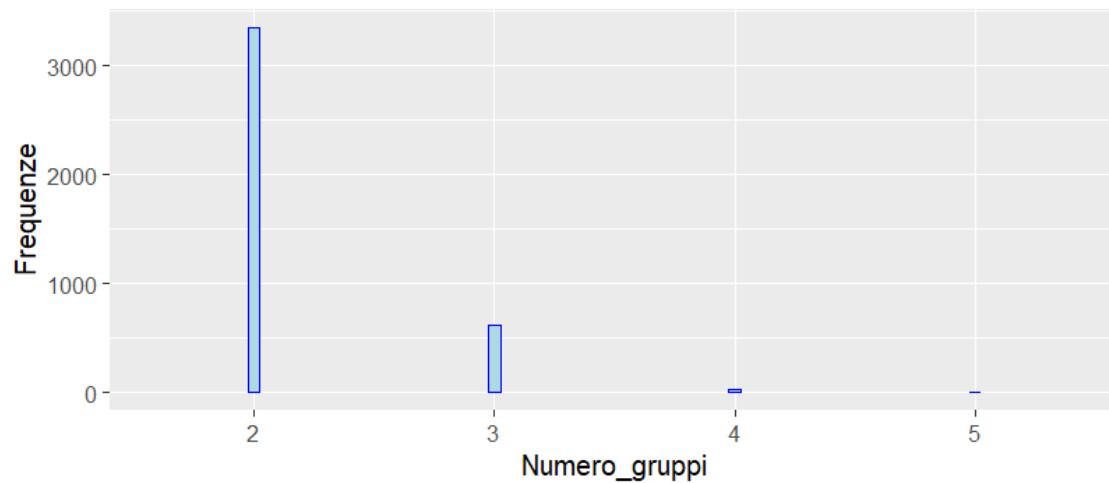


Figura 3.6: Distribuzione a posteriori del numero di gruppi, utilizzando la distribuzione a priori Stirling-Gamma di parametri $\alpha = 4$ e $\beta = 2$ sulla rete costituita da due gruppi.

	WAIC
Dirichlet-Multinomiae	644.65
Processo di Dirichlet	644.69
Pitman-Yor	644.61
Gnedin	644.81
Stirling-Gamma	644.89

Tabella 3.1: WAIC calcolato sui campioni \mathbf{z} ottenuti dalla rete costituita da due gruppi, impiegando le diverse distribuzioni a priori di tipo Gibbs con iperparametri che restituiscono 2 come valore atteso del numero di gruppi

fin'ora tutte mostrato risultati tutto sommato equivalenti , possiamo calcolare il criterio d'informazione WAIC per ognuna di esse e vedere quale di queste restituisce un valore più basso. Dalla Tabella 3.1 possiamo notare come in questo contesto la distribuzione a priori che restituisce WAIC minore sia quella di Pitman-Yor, mentre quella con valore più alto è quella di Stirling-Gamma. Le differenze nei valori di WAIC comunque, sono molto piccole e potrebbero rappresentare l'errore dovuto al campionamento Monte Carlo, piuttosto che una differenza nella performance

Andiamo adesso a vedere come cambiano i risultati appena visti, quando impostiamo gli iperparametri delle nostre cinque distribuzioni a priori in modo tale che il valore atteso del numero di gruppi risulti 5. In particolare per la distribuzione a priori Dirichlet-Multinomiale poniamo $\bar{H} = 9$ e $\sigma = -2/9$, per il Processo di Dirichlet mettiamo $\alpha = 1.2$, per la distribuzione a priori Pitman-Yor impostiamo $\sigma = 0.25$ e $\alpha = 0.35$, per Gnedin $\gamma = 0.55$ e per Stirling-Gamma $\alpha = 10$ e $\beta = 2$. Troviamo, anche in questo caso, la rappresentazione della distribuzione a posteriori del numero di gruppi per ciascuna distribuzione a priori nei grafici raffigurati in Figura 3.7, Figura 3.8, Figura 3.9, Figura 3.10 e Figura 3.11. Adesso che il valore atteso del numero di gruppi è superiore al loro valore effettivo, la sua distribuzione a posteriori diviene più variabile per le diverse distribuzioni a priori. In particolare, notiamo che per la Dirichlet-Multinomial la moda della distribuzione si sposta sul valore 3, mentre per le altre distribuzioni a priori rimane sul 2. La distribuzione meno variabile risulta quella ottenuta utilizzando la distribuzione a priori di Gnedin, per cui i valori del numero di gruppi vanno da 2 a 4, mentre quella che mostra maggiore variabilità, con un numero di gruppi che va da 2 a 6, è ottenuta grazie alla distribuzione a priori di Pitman-Yor. La partizione ottima \mathbf{z}^* , calcolata sempre grazie all'algoritmo SALSO con funzione di perdita Binder di parametro $a=1$, per le distribuzioni a priori Dirichlet-Multinomiale, Processo di Dirichlet, Pitman-Yor, Gnedin e Stirling-Gamma coincide con la partizione vera.

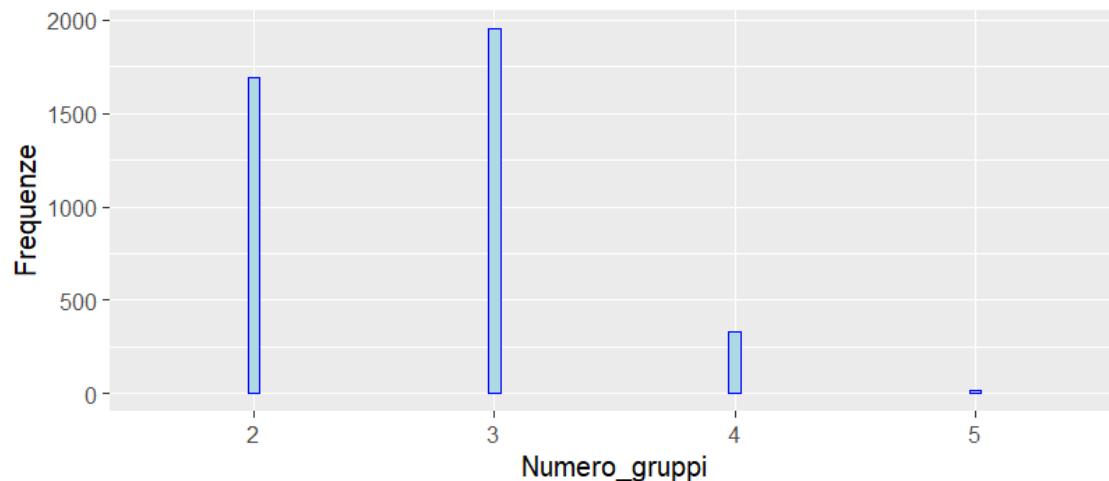


Figura 3.7: Distribuzione a posteriori del numero di gruppi, utilizzando la distribuzione a priori Dirichlet-Multinomiale di parametri $\bar{H} = 9$ e $\sigma = -2/9$ sulla rete costituita da due gruppi.

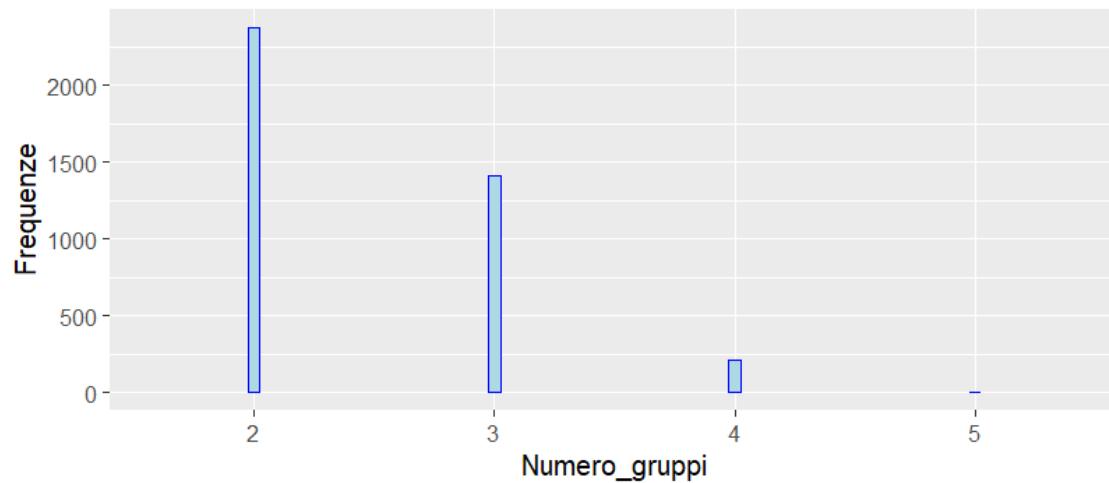


Figura 3.8: Distribuzione a posteriori del numero di gruppi, utilizzando la distribuzione a priori Processo di Dirichlet di parametro $\alpha = 1.2$ sulla rete costituita da due gruppi.

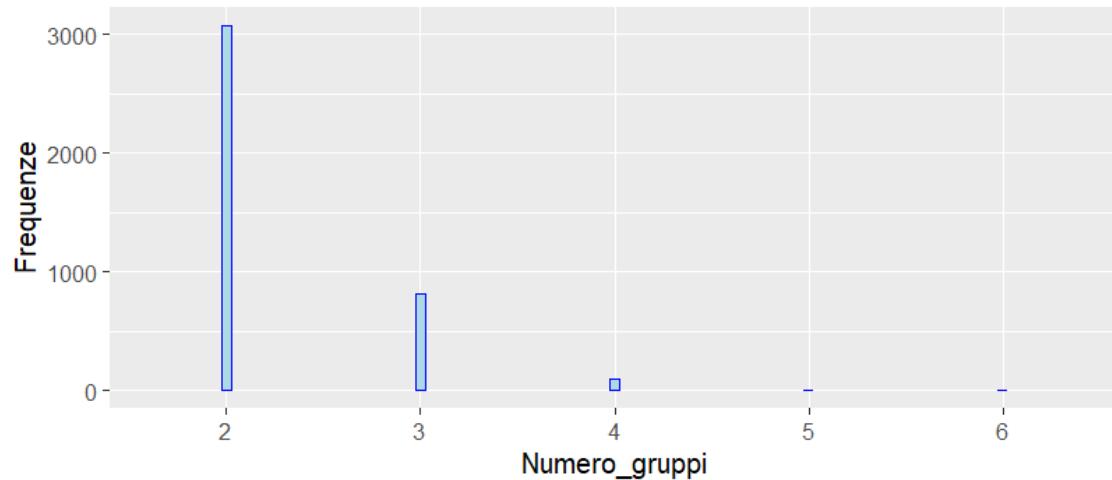


Figura 3.9: Distribuzione a posteriori del numero di gruppi, utilizzando la distribuzione a priori Pitman-Yor di parametri $\sigma = 0.25$ e $\alpha = 0.35$ sulla rete costituita da due gruppi.

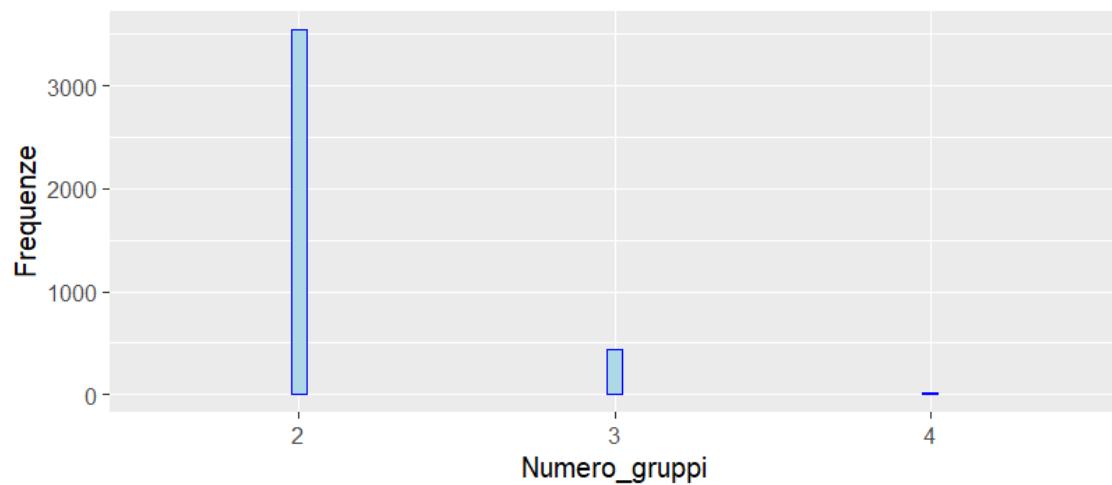


Figura 3.10: Distribuzione a posteriori del numero di gruppi, utilizzando la distribuzione a priori di Gnedin di parametro $\gamma = 0.55$ sulla rete costituita da due gruppi.

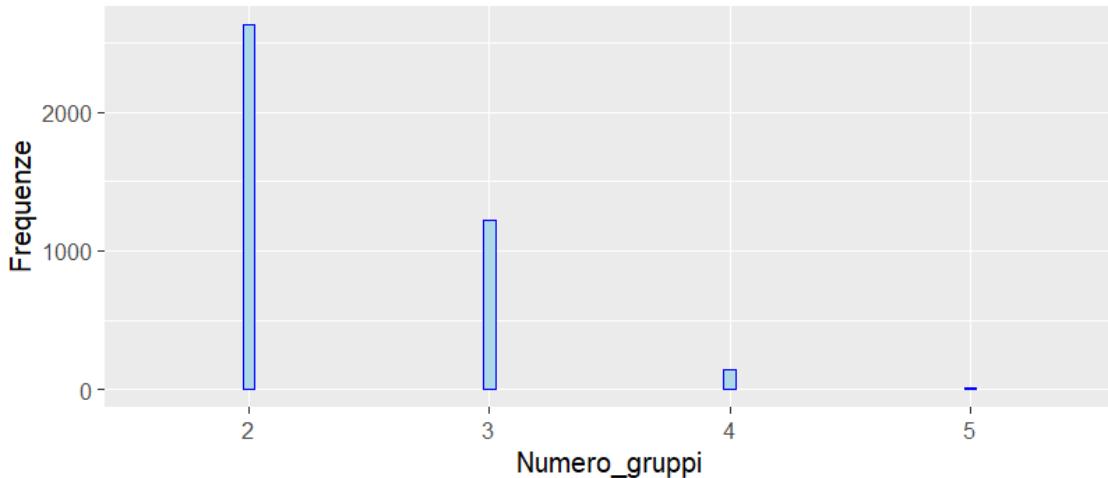


Figura 3.11: Distribuzione a posteriore del numero di gruppi, utilizzando la distribuzione a priori Stirling-Gamma di parametri $\alpha = 25$ e $\beta = 5$ sulla rete costituita da due gruppi.

	WAIC
Dirichlet-Multinomiae	644.73
Processo di Dirichlet	645.18
Pitman-Yor	645.23
Gnedin	644.58
Stirling-Gamma	644.97

Tabella 3.2: WAIC calcolato sui campioni \mathbf{z} ottenuti dalla rete costituita da due gruppi, impiegando le diverse distribuzioni a priori di tipo Gibbs con iperparametri che restituiscono 5 come valore atteso del numero di gruppi

Confrontiamo la performance delle diverse distribuzioni a priori sempre attraverso il criterio d'informazione WAIC, i cui valori sono rappresentati nella Tabella 3.2. Anche in questo contesto i valori di WAIC non sono troppo diversi tra di loro, in particolare è la distribuzione a priori di Pitman-Yor quella che produce il modello con la performance peggiore, in quanto ha il valore di WAIC più alto, probabilmente a causa della maggiore variabilità nella distribuzione a posteriore del numero di gruppi. La distribuzione a priori di Gnedin risulta invece produrre il modello migliore.

La seconda rete simulata è composta da $n = 90$ nodi e $H_0 = 3$ gruppi, dove i primi 25 appartengono al primo gruppo, i seguenti 30 al secondo e i restanti 35 al terzo. La matrice a blocchi Θ ha sulla diagonale principale tutti valori pari a 0.8, probabilità che due nodi appartenenti allo stesso gruppo abbiano un legame e 0.05 al di fuori, probabilità che due nodi appartenenti a due gruppi differenti abbiano in

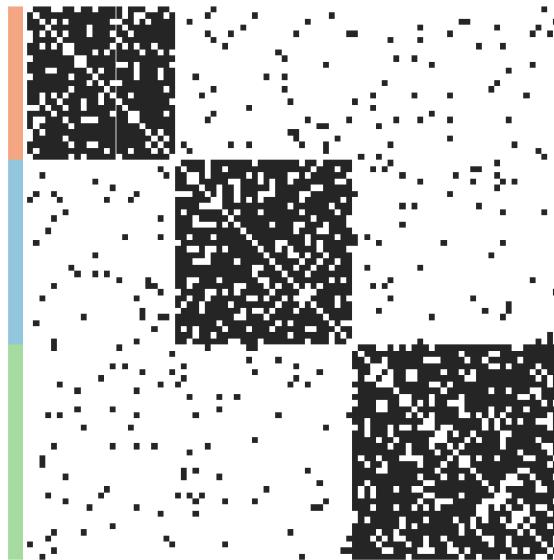


Figura 3.12: Matrice di adiacenza per la rete composta da tre gruppi

legame :

$$\Theta = \begin{bmatrix} 0.8 & 0.05 & 0.05 \\ 0.05 & 0.8 & 0.05 \\ 0.05 & 0.05 & 0.8 \end{bmatrix}$$

Dalla simulazione della matrice di adiacenza \mathbf{Y} utilizzando queste informazioni, otteniamo la Figura 3.12. I tre gruppi risultano ancora ben distinti tra loro, nonostante abbiamo diminuito la probabilità che due nodi appartenenti allo stesso gruppo abbiano un legame. In primo luogo andiamo ad implementare i modelli per blocchi stocastici estesi impostando gli iperparametri delle diverse priori in modo che il valore atteso del numero di gruppi sia 3, come quello effettivo. In particolare impostiamo $\bar{H} = 10$ e $\sigma = -0.06$ per la distribuzione a priori Dirichlet-Multinomiale, $\alpha = 0.45$ per il Processo di Dirichlet, $\sigma = 0.175$ e $\alpha = 0.1$ per la distribuzione a priori Pitman-Yor, $\gamma = 0.75$ per Gneden e $\alpha = 6$ e $\beta = 2$ per Stirling-Gamma. Per il resto delle impostazioni necessarie per eseguire l’Algoritmo 1, impieghiamo quelle descritte per la rete costituita da due gruppi. Andiamo a vedere come si distribuisce a posteriori il numero di gruppi a seconda dalla distribuzione a priori impiegata, attraverso i grafici rappresentati in Figura 3.13, Figura 3.14, Figura 3.15, Figura 3.16 e Figura 3.17.

Possiamo notare che per tutte le *prior* la distribuzione a posteriori del numero di gruppi si concentra sul valore 3, effettivo numero di gruppi presenti nella rete. La distribuzione che mostra minore variabilità quella prodotta dalla distribuzione a priori di Gneden, dove il numero di gruppi varia solo tra 3 e 4, mentre per tutte le altre arriva anche a 5. Calcoliamo la partizione ottima \mathbf{z}^* grazie all’algoritmo

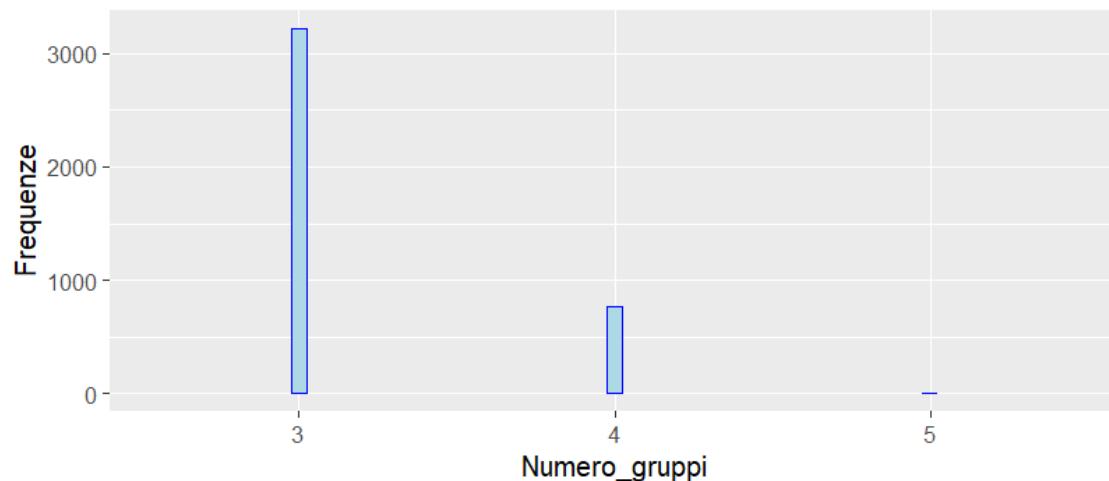


Figura 3.13: Distribuzione a posteriori del numero di gruppi, utilizzando la distribuzione a priori Dirichlet-Multinomiale di parametri $\bar{H} = 10$ e $\sigma = -0.06$ sulla rete costituita da tre gruppi.

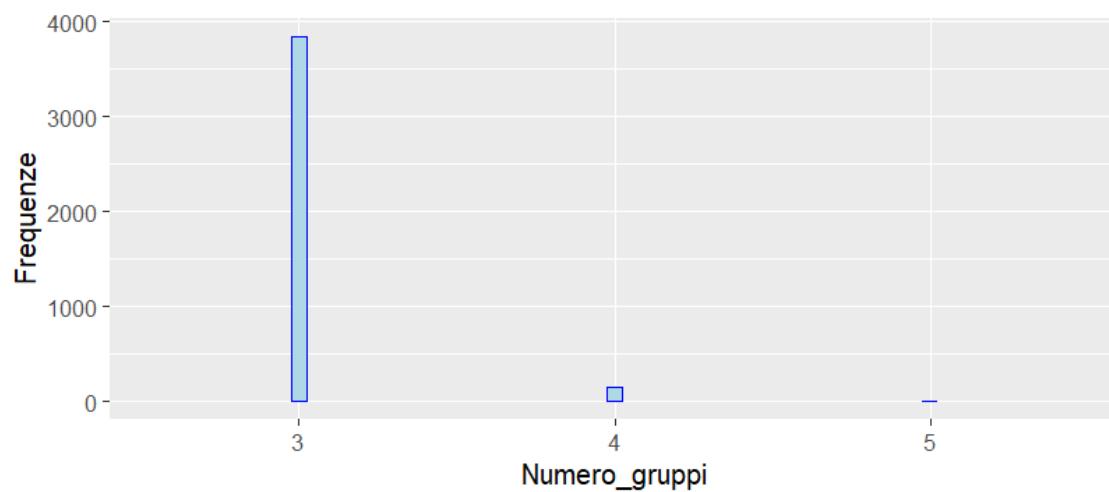


Figura 3.14: Distribuzione a posteriori del numero di gruppi, utilizzando la distribuzione a priori Processo di Dirichlet di parametro $\alpha = 0.45$ sulla rete costituita da tre gruppi.

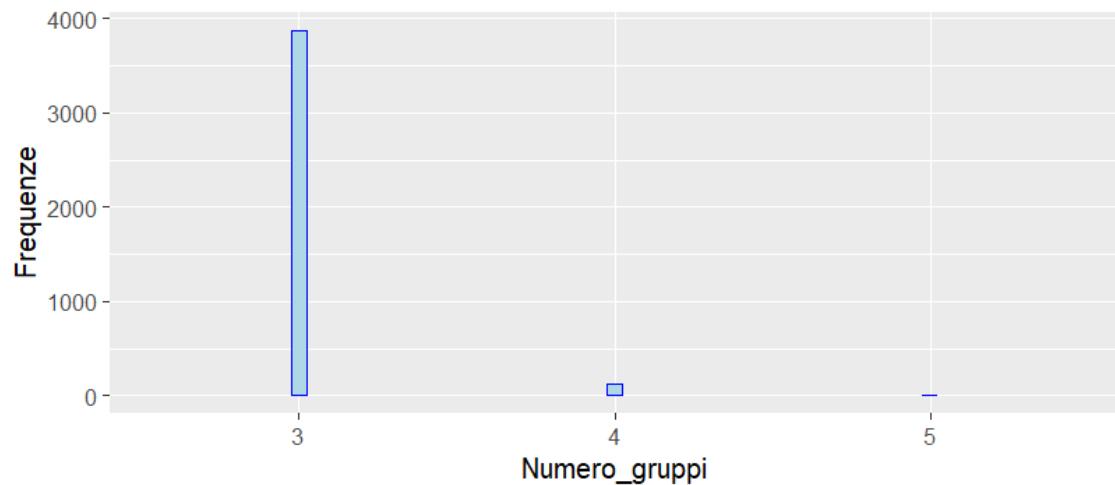


Figura 3.15: Distribuzione a posteriore del numero di gruppi, utilizzando la distribuzione a priori Pitman-Yor di parametri $\sigma = 0.175$ e $\alpha = 0.1$ sulla rete costituita da tre gruppi.

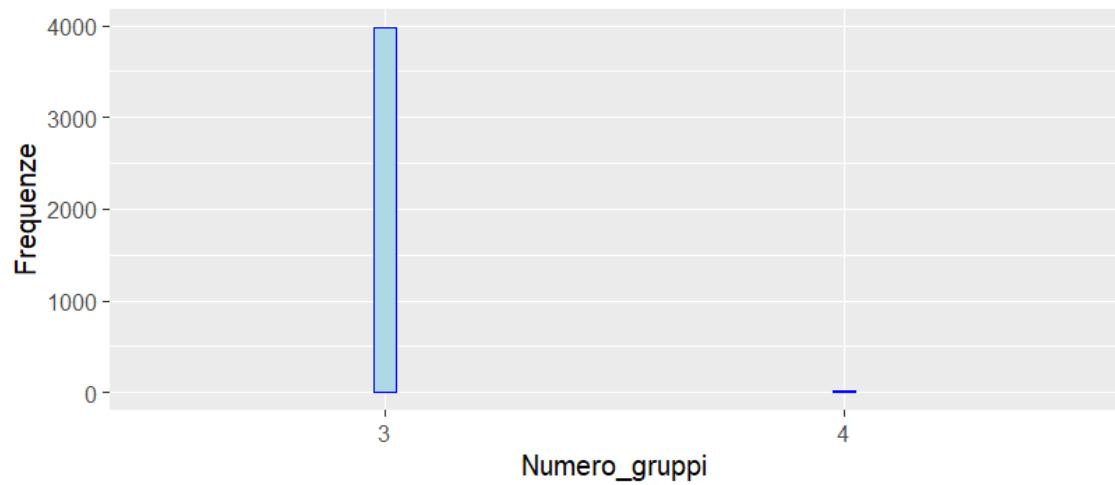


Figura 3.16: Distribuzione a posteriore del numero di gruppi, utilizzando la distribuzione a priori di Gnedin di parametro $\gamma = 0.75$ sulla rete costituita da tre gruppi.

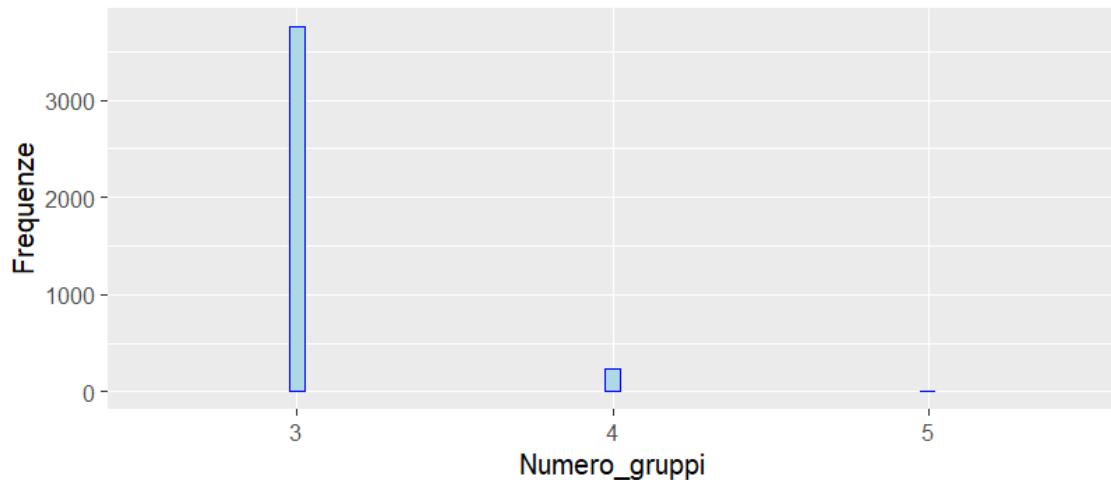


Figura 3.17: Distribuzione a posteriore del numero di gruppi, utilizzando la distribuzione a priori Stirling-Gamma di parametri $\alpha = 6$ e $\beta = 2$ sulla rete costituita da tre gruppi.

	WAIC
Dirichlet-Multinomiae	2415.9
Processo di Dirichlet	2415.54
Pitman-Yor	2414.75
Gnedin	2415.57
Stirling-Gamma	2415.42

Tabella 3.3: WAIC calcolato sui campioni \mathbf{z} ottenuti dalla rete costituita da tre gruppi, impiegando le diverse distribuzioni a priori di tipo Gibbs con iperparametri che restituiscono 3 come valore atteso del numero di gruppi

SALSO con funzione di perdita di Binder di parametro $a = 1$. Questa coincide con la partizione vera per tutte le distribuzioni a priori. Nonostante la diminuzione effettuata nella probabilità che due nodi appartenenti allo stesso gruppo abbiano un legame, i tre gruppi rimangono ben distinti tra loro. Per confrontare la performance delle varie distribuzioni a priori, ricorriamo sempre al WAIC, i cui valori sono indicati nella Tabella 3.3. Il modello che presenta valore del WAIC minore è quello ottenuto dalla distribuzione a priori Stirling-Gamma, mentre quella con valore più alto è quello che utilizza la Dirichlet-Multinomiale, seppur le differenze rimangono minime e quindi attribuibile all'errore Monte Carlo.

In seguito, vediamo come cambiano questi risultati se impostiamo gli iperparametri delle distribuzioni a priori in modo che il valore atteso del numero di gruppi sia 9. Ovvero impostiamo $\bar{H} = 17$ e $\sigma = -3.8/17$ per la distribuzione a priori Dirichlet-Multinomiale, $\alpha = 2.3$ per il processo di Dirichlet, $\sigma = 0.315$ e $\alpha = 0.6$ per la distribuzione a priori Pitman-Yor, $\gamma = 0.475$ per Gnedin e $\alpha = 18$ e $\beta = 2$ per

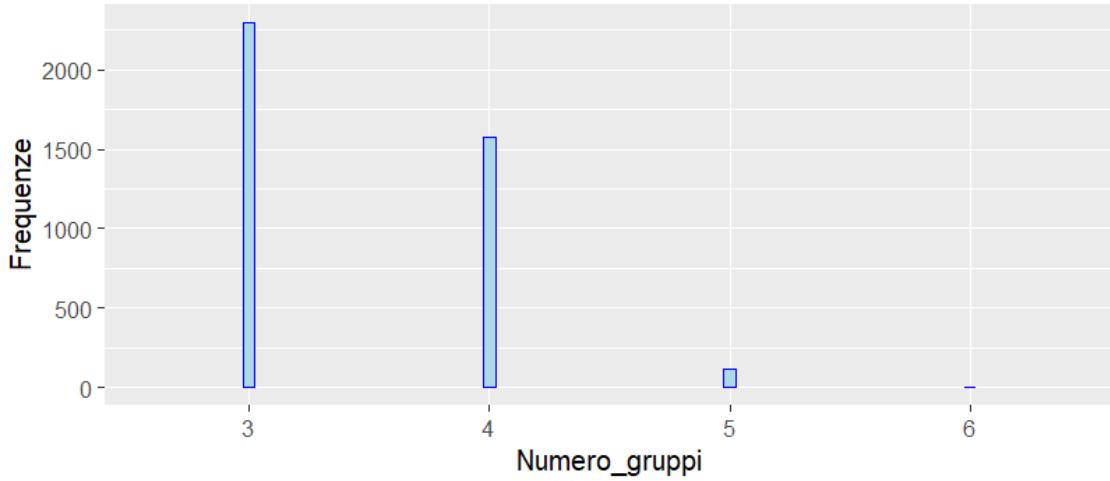


Figura 3.18: Distribuzione a posteriore del numero di gruppi, utilizzando la distribuzione a priori Dirichlet-Multinomiale di parametri $\bar{H} = 17$ e $\sigma = -3.8/17$ sulla rete costituita da tre gruppi.

	WAIC
Dirichlet-Multinomiae	2415.91
Processo di Dirichlet	2416.03
Pitman-Yor	2415.97
Gnedin	2415.59
Stirling-Gamma	2416.03

Tabella 3.4: WAIC calcolato sui campioni \mathbf{z} ottenuti dalla rete costituita da tre gruppi, impiegando le diverse distribuzioni a priori di tipo Gibbs con iperparametri che restituiscono 9 come valore atteso del numero di gruppi

Stirling-Gamma. Rappresentiamo le nuove distribuzioni a posteriori del numero di gruppi attraverso grafici in Figura 3.18, Figura 3.19, Figura 3.20, Figura 3.21, Figura 3.22. Osserviamo che le distribuzioni sono più variabili rispetto alle precedenti per via dei nuovi iperparametri, ma tutte le distribuzioni mantengono come valore della moda 3. Le distribuzioni ottenute grazie alla distribuzione a priori di Gnedin e Stirling-Gamma sono quelle che variano meno, tra 3 e 4, mentre le altre arrivano anche a 6. Coerentemente con quanto osservato nei grafici le partizioni ottime \mathbf{z}^* calcolate con l'algoritmo SALSO con funzione di perdita Binder di parametro $a = 1$, coincidono con la partizione vera per tutte le distribuzioni a priori. Confrontiamo infine le performance dei modelli sotto le varie distribuzioni a priori attraverso il WAIC, per vedere quale produce il modelli migliore. I valori di questo criterio d'informazione sono rappresentati nella Tabella 3.4. Notiamo che i valori di WAIC rimangono ancora una volta molto vicini tra di loro. I modelli prodotti dalle distribuzioni a priori

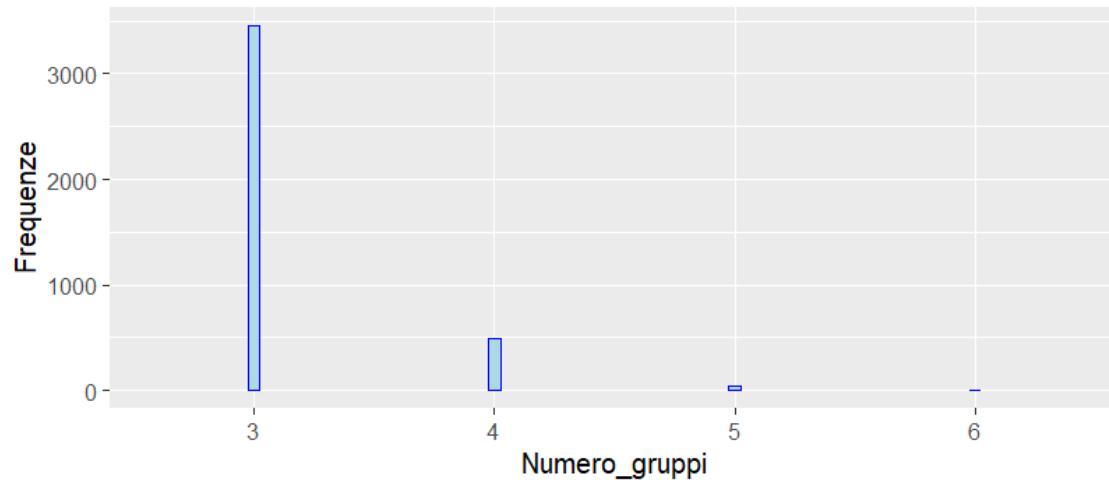


Figura 3.19: Distribuzione a posteriore del numero di gruppi, utilizzando la distribuzione a priori Processo di Dirichlet di parametro $\alpha = 2.3$ sulla rete costituita da tre gruppi.

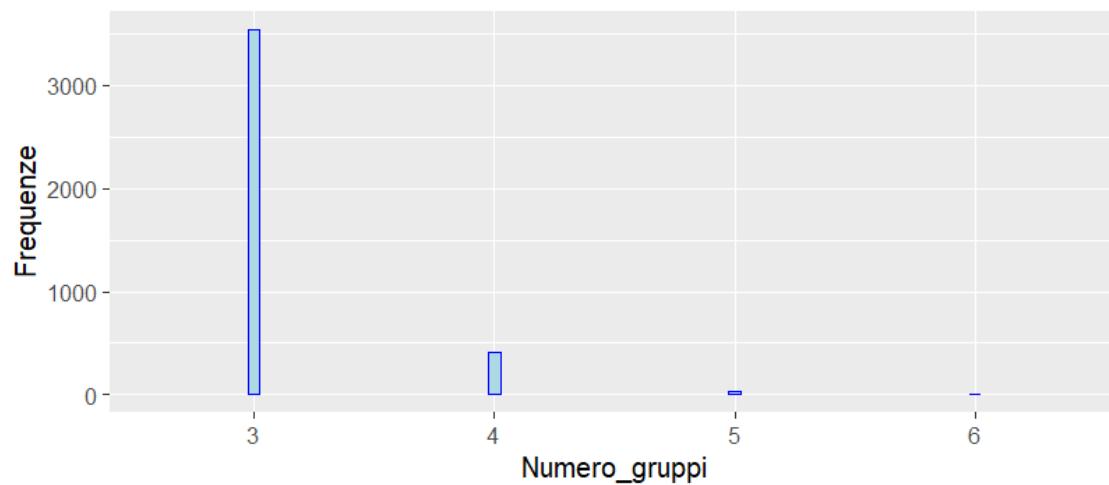


Figura 3.20: Distribuzione a posteriore del numero di gruppi, utilizzando la distribuzione a priori Pitman-Yor di parametri $\sigma = 0.315$ e $\alpha = 0.6$ sulla rete costituita da tre gruppi.

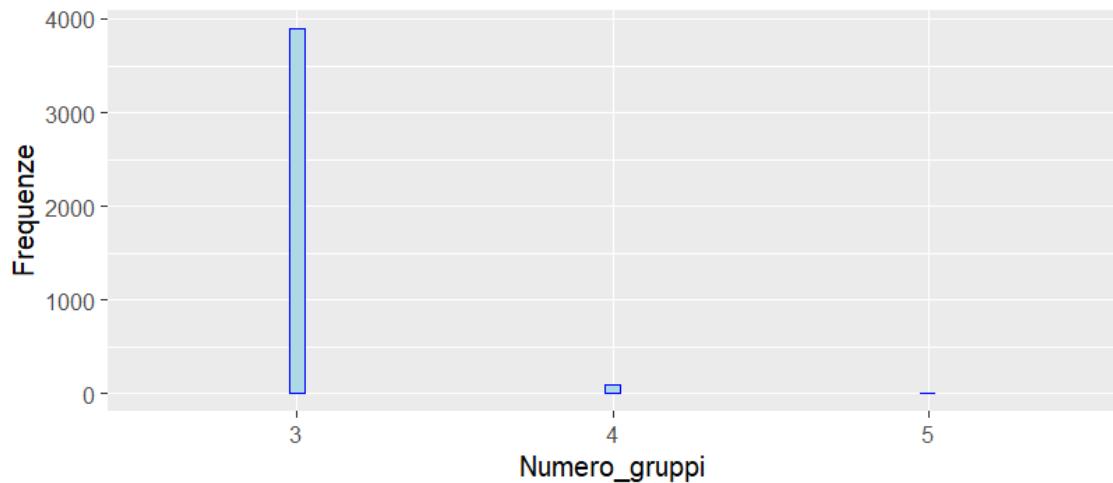


Figura 3.21: Distribuzione a posteriore del numero di gruppi, utilizzando la distribuzione a priori di Gnedin di parametro $\gamma = 0.475$ sulla rete costituita da tre gruppi.

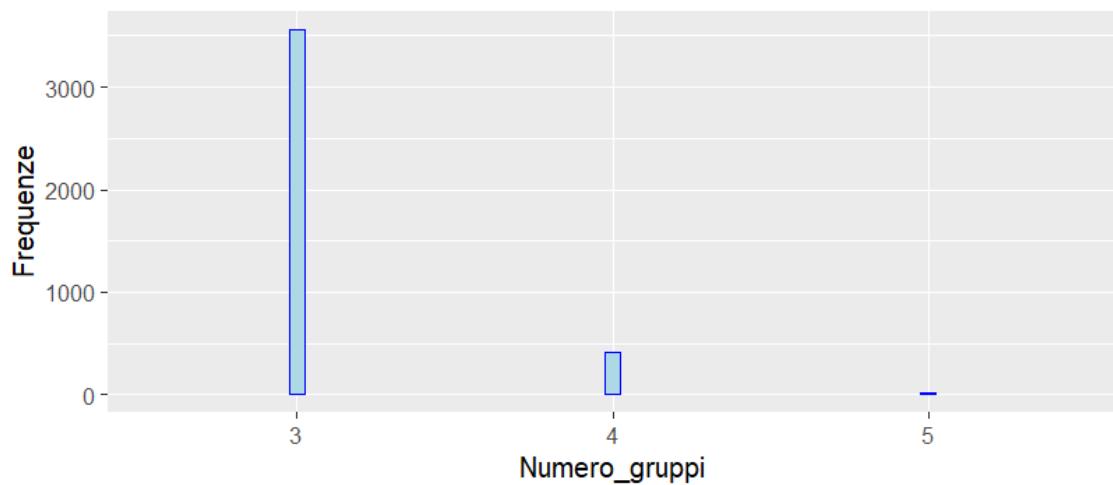


Figura 3.22: Distribuzione a posteriore del numero di gruppi, utilizzando la distribuzione a priori Stirling-Gamma di parametri $\alpha = 18$ e $\beta = 2$ sulla rete costituita da tre gruppi.

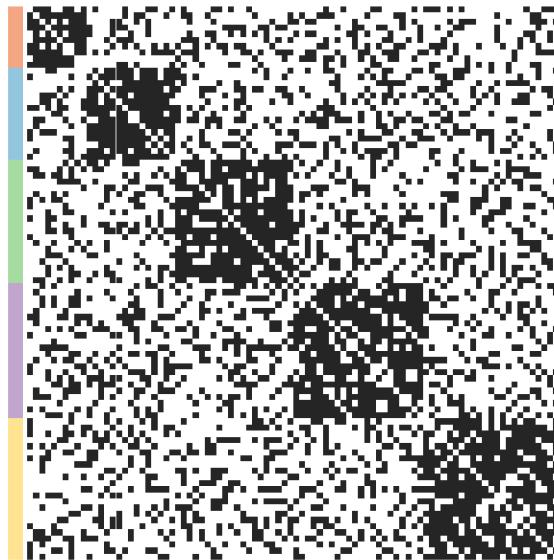


Figura 3.23: Matrice di adiacenza per la rete composta da cinque gruppi

Processo di Dirichlet e Stirling-Gamma hanno un valore del WAIC maggiore, mentre il modello con valore inferiore risulta essere quello ottenuto dalla distribuzione a priori di Gnedin.

Simuliamo una terza rete sempre di $n = 90$ nodi, aumentando però il numero di gruppi a $H_0 = 5$, in particolare i primi 10 nodi appartengono al primo gruppo i seguenti 15 al secondo, i 20 successivi al terzo, i 22 dopo al quarto e i restanti 23 al quinto. Cambiamo inoltre le probabilità contenuta nella matrice a blocchi Θ , lasciando la probabilità che due nodi appartenenti allo stesso gruppo abbiano un legame pari a 0.8 e aumentando la probabilità che due nodi appartenenti a due gruppi diversi abbiano un legame a 0.3:

$$\Theta = \begin{bmatrix} 0.8 & 0.3 & 0.3 & 0.3 & 0.3 \\ 0.3 & 0.8 & 0.3 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0.8 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0.3 & 0.8 & 0.3 \\ 0.3 & 0.3 & 0.3 & 0.3 & 0.8 \end{bmatrix}$$

La matrice di adiacenza \mathbf{Y} che otteniamo è rappresentata in Figura 3.23. I gruppi non sono più ben distinti come nelle matrici viste in precedenza, andiamo a vedere come questo si ripercuote sui nostri risultati. Di nuovo come prima combinazione degli iperparametri delle nostre priori ne scegliamo una che restituisca come valore atteso del numero di gruppi il vero valore, ovvero 5. Ne risulta che la distribuzione a priori Dirichlet-Multinomiale ha come iperparametri $\bar{H} = 16$ e $\sigma = -1.25/16$, il

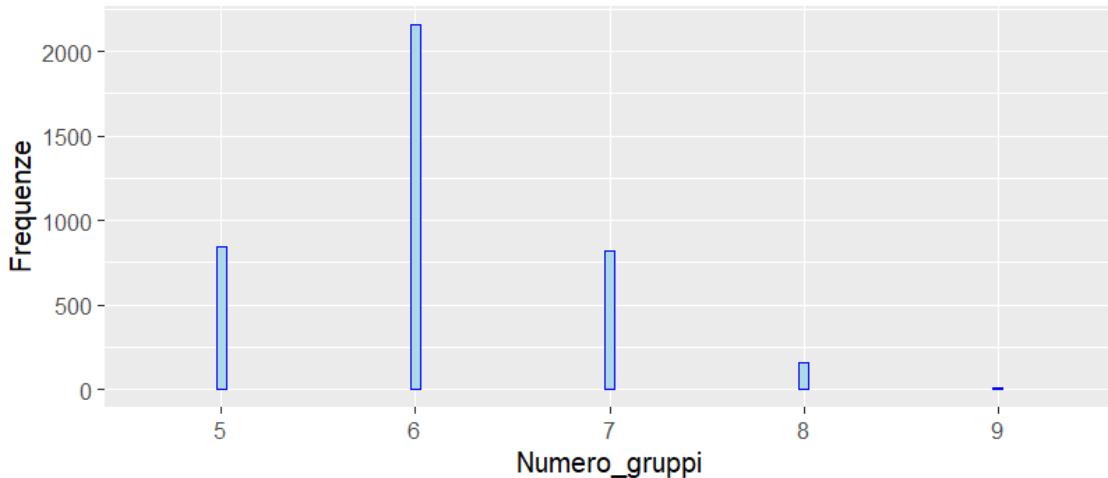


Figura 3.24: Distribuzione a posteriori del numero di gruppi, utilizzando la distribuzione a priori Dirichlet-Multinomiale di parametri $\bar{H} = 16$ e $\sigma = -1.25/16$ sulla rete costituita da cinque gruppi.

Processo di Dirichlet ha $\alpha = 0.98$, la distribuzione a priori Pitman-Yor presenta $\sigma = 0.225$ e $\alpha = 0.3$, per Gneden $\gamma = 0.6$ e la Stirling-Gamma ha $\alpha = 10$ e $\beta = 2$. Ancora una volta per l’implementazione dell’Algoritmo 1 ci rifacciamo alle impostazioni elencate nel primo esempio. Andiamo ad analizzare come si comportano questi modelli sulla nuova rete, andando prima di tutto ad osservare la distribuzione a posteriori del numero di gruppi per le diverse distribuzioni a priori impiegate, illustrata nei grafici in Figura 3.24, Figura 3.25, Figura 3.26, Figura 3.27 e Figura 3.28. Notiamo subito come i cambiamenti apportati alla matrice a blocchi abbiano notevolmente aumentato la variabilità delle distribuzioni a posteriori del numero di gruppi, andando da 5 fino anche a 11 per le distribuzioni a priori Processo di Dirichlet, Pitman-Yor e Stirling-Gamma. Inoltre possiamo osservare che per tutte le distribuzioni a priori impiegate la moda si trova sul valore 6 e non su 5, effettivo numero di gruppi presente nella rete. In seguito osserviamo che le partizioni ottime \mathbf{z}^* calcolate con l’algoritmo SALSO con funzione di perdita di Binder di parametro $a = 1$ non coincidono più con la partizione vera, questo è coerente con il comportamento della distribuzione a posteriori del numero di gruppi appena descritto. In particolare nessuna delle partizioni trovate riesce a classificare correttamente i nodi 19 e 61. Nel caso delle partizioni ottenute con la distribuzione a priori Dirichlet-Multinomiale, Pitman-Yor e Gneden, questi due nodi vanno a costituire due ulteriori gruppi distinti, mentre per le altre vanno a formare insieme un ulteriore gruppo. Se cambiamo il parametro della funzione di perdita di Binder a $a = 1.7$, così da controllare il numero di gruppi, la partizione ottenuta corrisponde a quella effettiva per tutte le

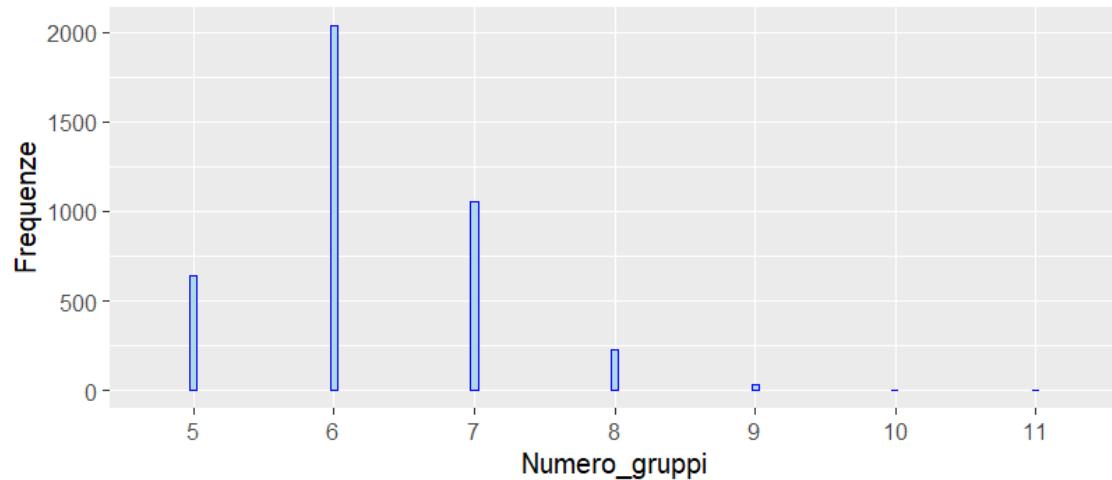


Figura 3.25: Distribuzione a posteriore del numero di gruppi, utilizzando la distribuzione a priori Processo di Dirichlet di parametro $\alpha = 0.98$ sulla rete costituita da cinque gruppi

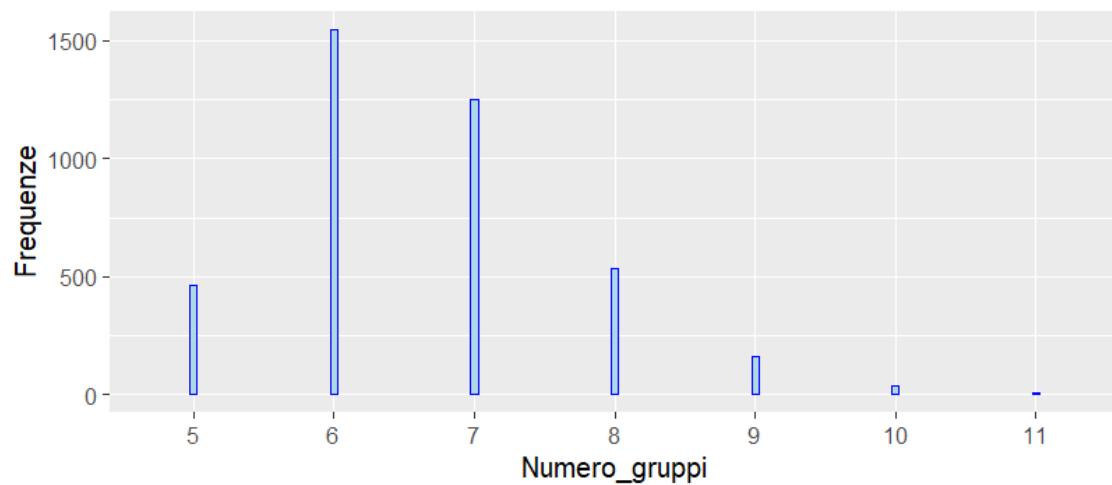


Figura 3.26: Distribuzione a posteriore del numero di gruppi, utilizzando la distribuzione a priori Pitman-Yor di parametri $\sigma = 0.225$ e $\alpha = 0.3$ sulla rete costituita da cinque gruppi.

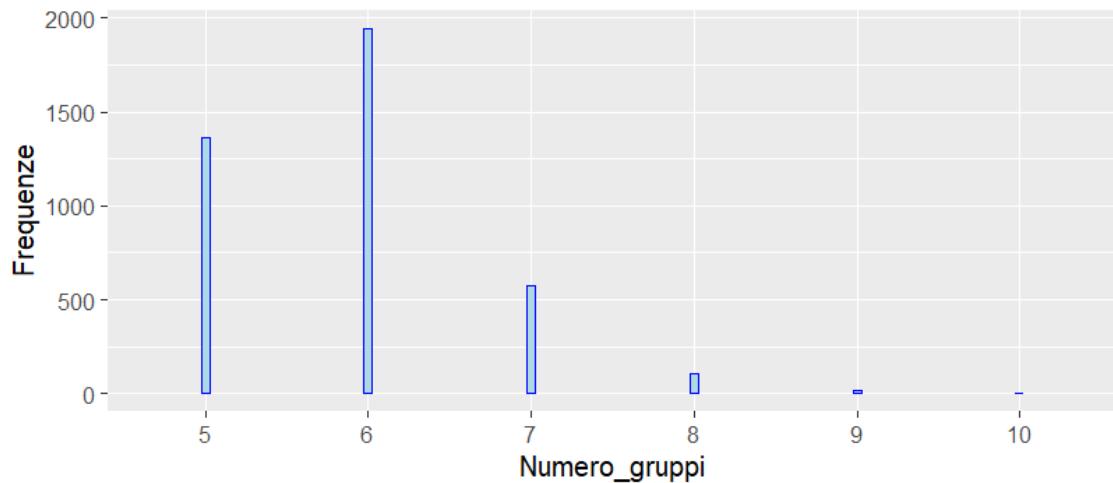


Figura 3.27: Distribuzione a posteriore del numero di gruppi, utilizzando la distribuzione a priori di Gneden di parametro $\gamma = 0.6$ sulla rete costituita da cinque gruppi.

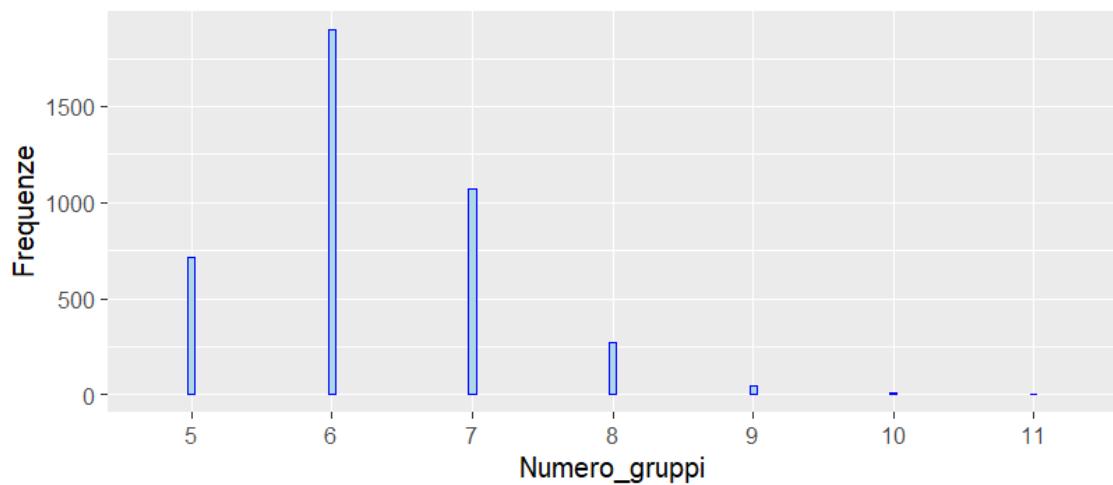


Figura 3.28: Distribuzione a posteriore del numero di gruppi, utilizzando la distribuzione a priori Stirling-Gamma di parametri $\alpha = 10$ e $\beta = 2$ sulla rete costituita da cinque gruppi.

	WAIC
Dirichlet-Multinomilae	4765.89
Processo di Dirichlet	4763.92
Pitman-Yor	4767.07
Gnedin	4765.38
Stirling-Gamma	4766.04

Tabella 3.5: WAIC calcolato sui campioni \mathbf{z} ottenuti dalla rete costituita da cinque gruppi, impiegando le diverse distribuzione a priori di tipo Gibbs con iperparametri che restituiscono 5 come valore atteso del numero di gruppi

distribuzioni a priori. Infine confrontiamo le performance dei modelli prodotti dalle diverse distribuzioni a priori sempre attraverso l'indice WAIC, i cui valori si trovano nella Tabella 3.5. Osserviamo che il modello che produce WAIC minore è quello che utilizza la distribuzione a priori Processo di Dirichlet, mentre quello con valore più alto ricorre alla Pitman-Yor. In generale possiamo notare che i modelli che impiegano le distribuzioni a priori che individuano come partizione ottima, nella prima opzione di parametro per la funzione di perdita, quella composta da 6 gruppi hanno un valore del WAIC inferiore rispetto a quelli che utilizzano distribuzioni a priori che producono una partizione ottima, nelle stesse condizioni, costituita da 7 gruppi.

Vediamo ora i nuovi i risultati, impostando gli iperparametri delle distribuzioni a priori in modo che il numero atteso di gruppi sia 10. In particolare abbiamo la Dirichlet-Multinomiale con $\bar{H} = 19$ e $\sigma = -4.5/19$, il Processo di Dirichlet con $\alpha = 2.7$, Pitman-Yor con $\sigma = 0.3$ e $\alpha = 0.825$, Gnedin con $\gamma = 0 - 46$ e la Stirling-Gamma con $\alpha = 20$ e $\beta = 2$. Partiamo sempre con il visualizzare la distribuzione a posteriori del numero di gruppi per le diverse distribuzioni a priori, queste le troviamo nei grafici rappresentati in Figura 3.29, Figura 3.30, Figura 3.31, Figura 3.32 e Figura 3.33. Osserviamo come la variabilità sia ulteriormente aumentata, nello specifico la distribuzione a posteriori per il numero di gruppi che mostra range più ampio è quella ottenuta dalla distribuzione a priori di Pitman-Yor, per la quale il valore massimo è 14. Inoltre tutte le distribuzioni hanno la moda sul valore 7, tranne quella derivata grazie alla distribuzione a priori di Gnedin, per la quale il valore centrale rimane 6. Per quanto riguarda la partizione ottima \mathbf{z}^* , trovata utilizzando sempre l'algoritmo SALSO con funzione di perdita di Binder con parametro $a = 1$, invece, tutte le distribuzioni a priori forniscono risultati diversi. In particolare, la Stirling-Gamma è l'unica che produce una partizione ottima che coincide con quella vera, le distribuzioni a priori Dirichlet-Multinomiale e Gnedin

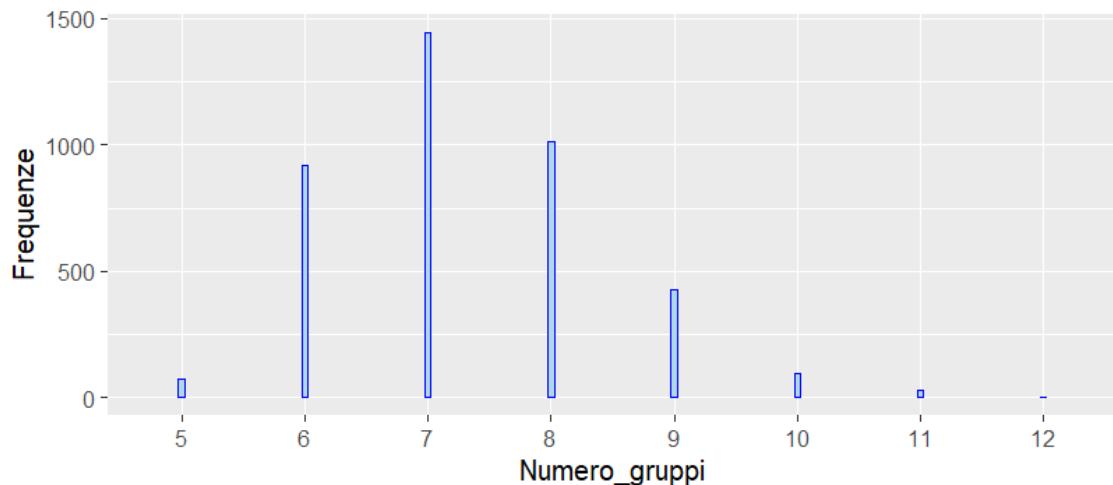


Figura 3.29: Distribuzione a posteriori del numero di gruppi, utilizzando la distribuzione a priori Dirichlet-Multinomiale di parametri $\bar{H} = 19$ e $\sigma = -4.5/19$ sulla rete costituita da cinque gruppi.

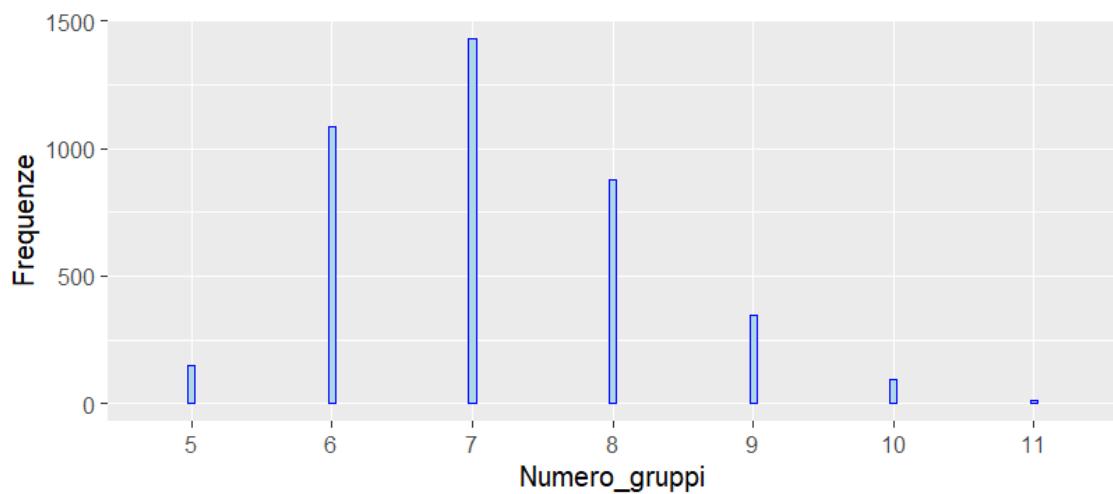


Figura 3.30: Distribuzione a posteriori del numero di gruppi, utilizzando la distribuzione a priori Processo di Dirichlet di parametro $\alpha = 2.7$ sulla rete costituita da cinque gruppi.

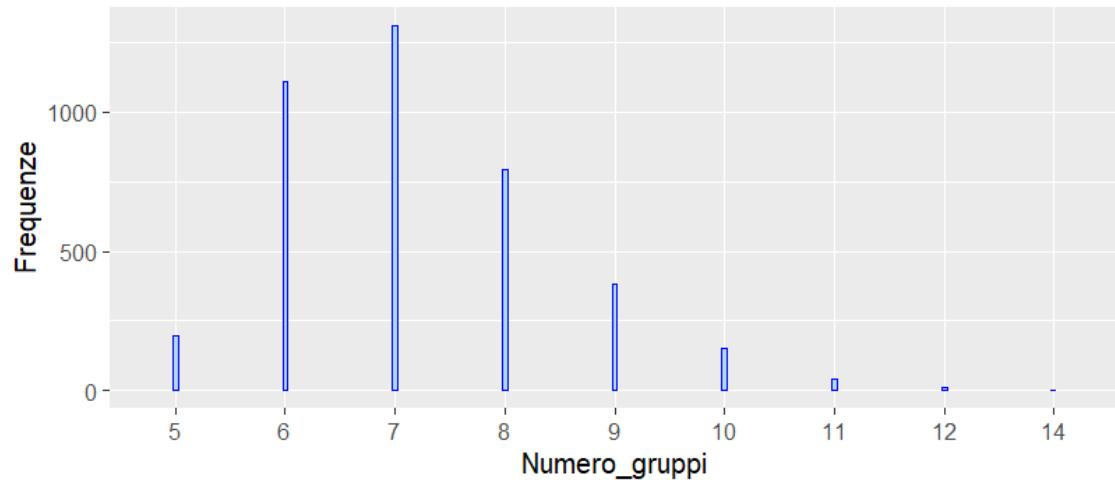


Figura 3.31: Distribuzione a posteriore del numero di gruppi, utilizzando la distribuzione a priori Pitman-Yor di parametri $\sigma = 0.3$ e $\alpha = 0.825$ sulla rete costituita da cinque gruppi.

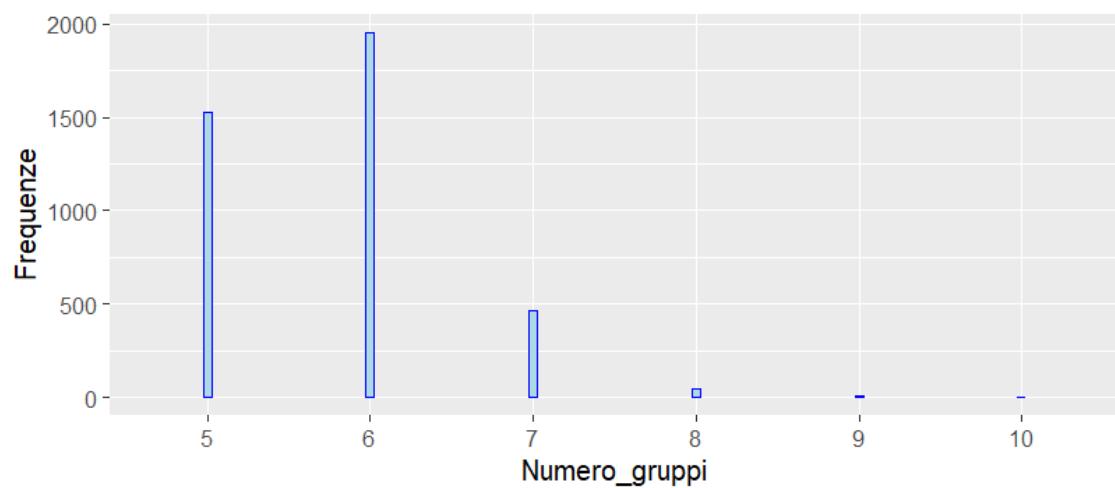


Figura 3.32: Distribuzione a posteriore del numero di gruppi, utilizzando la distribuzione a priori di Gnedin di parametro $\gamma = 0.46$ sulla rete costituita da cinque gruppi.

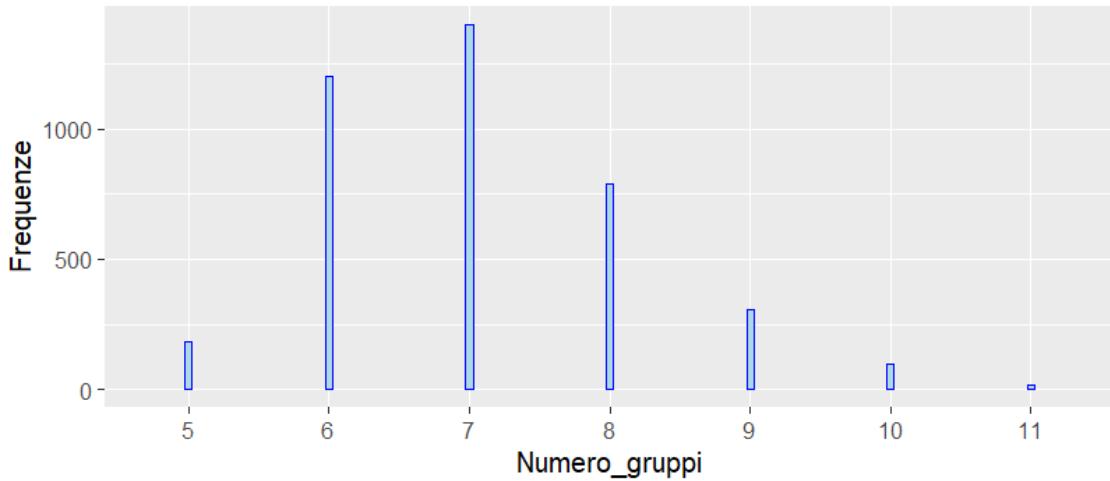


Figura 3.33: Distribuzione a posteriore del numero di gruppi, utilizzando la distribuzione a priori Stirling-Gamma di parametri $\alpha = 20$ e $\beta = 2$ sulla rete costituita da cinque gruppi.

	WAIC
Dirichlet-Multinomiae	4765.85
Processo di Dirichlet	4767.93
Pitman-Yor	4768.63
Gnedin	4766.64
Stirling-Gamma	4765.49

Tabella 3.6: WAIC calcolato sui campioni \mathbf{z} ottenuti dalla rete costituita da cinque gruppi, impiegando le diverse distribuzioni a priori di tipo Gibbs con iperparametri che restituiscono 10 come valore atteso del numero di gruppi

producono una partizione ottima composta da 6 gruppi, dove per la prima il sesto gruppo è costituito dal nodo 19, mentre per la seconda l’ulteriore gruppo contiene sia il nodo 19 che il 61. Le distribuzioni a priori Processo di Dirichlet e Pitman-Yor forniscono invece una partizione ottima costituita da 7 gruppi, per quest’ultima troviamo un gruppo costituito dal nodo 19 e un altro dal nodo 61, mentre per la prima un gruppo contiene i nodi 19 e 23 , mentre l’altro ancora una volta il nodo 61. Se cambiamo il parametro della funzione di perdita in $a = 1.7$, otteniamo una partizione ottima che coincide con quella vera per tutte le distribuzioni a priori tranne la Dirichlet-Multinomiale. Concludiamo l’analisi di questa rete ancora una volta confrontando la performance dei modelli prodotti dalle diverse distribuzioni a priori con il criterio d’informazione WAIC, i cui valori sono riportati nella Tabella 3.6. Notiamo che il modello che ricorre a distribuzione a priori Stirling-Gamma è quello con valore WAIC minore, questo è coerente con ciò che abbiamo osservato

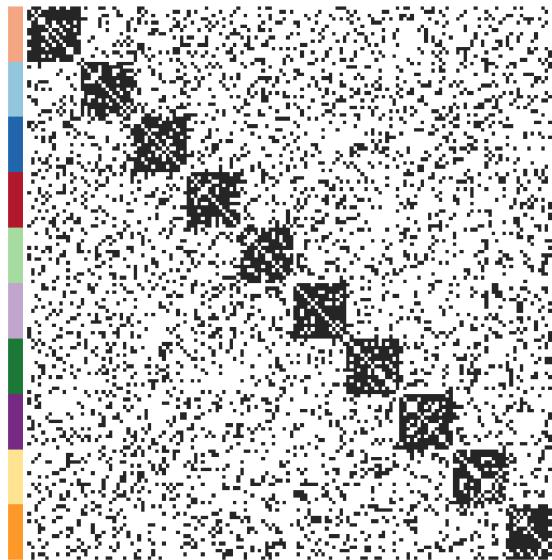


Figura 3.34: Matrice di adiacenza per la rete composta da dieci gruppi

durante l’analisi della partizione ottima. Mentre il modello con valore maggiore di WAIC è stato prodotto dalla distribuzione a priori di Pitman-Yor.

Concludiamo questa sezione con una rete composta da $n = 150$ nodi e $H_0 = 10$ gruppi, ognuno contenente 15 nodi. Abbiamo ancora una volta cambiato le probabilità contenute nella matrice a blocchi Θ , diminuendo la probabilità che due nodi appartenenti a due gruppi differenti abbiano un legame:

$$\Theta = \begin{bmatrix} 0.8 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.8 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.8 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.8 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.8 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.8 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.8 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.8 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.8 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.8 \end{bmatrix}$$

La matrice di adiacenza che otteniamo grazie a queste informazioni è rappresentata in Figura 3.34. Notiamo che i gruppi sono maggiormente distinti rispetto all’ultimo caso analizzato. Iniziamo anche in questo caso con l’analisi dei risultati ottenuti impostando gli iperparametri delle distribuzioni a priori in modo tale che il numero atteso di gruppi coincida con quello vero, ovvero 10. Per la distribuzione a priori

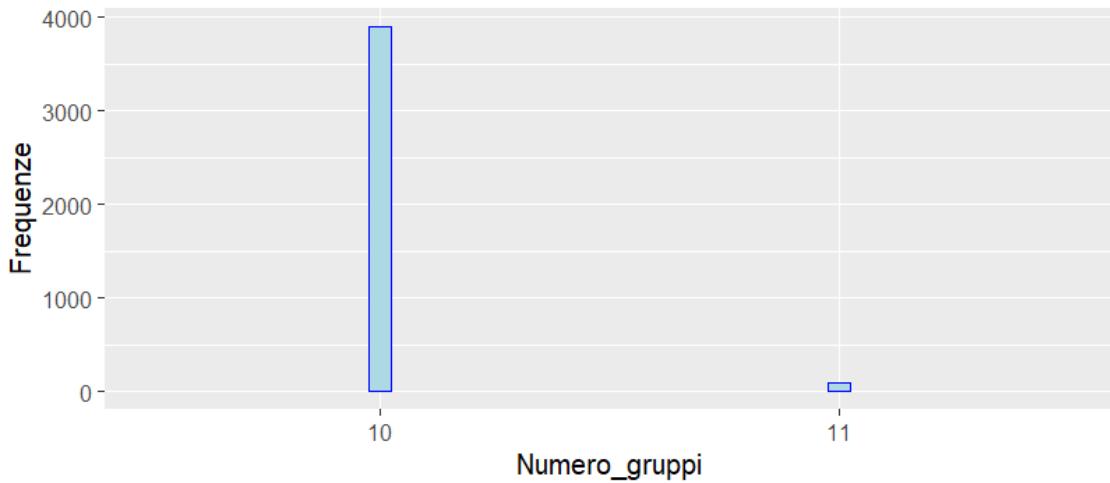


Figura 3.35: Distribuzione a posteriore del numero di gruppi, utilizzando la distribuzione a priori Dirichlet-Multinomiale di parametri $\bar{H} = 17$ e $\sigma = -4/17$ sulla rete costituita da dieci gruppi.

Dirichlet-Multinomiale otteniamo $\bar{H} = 17$ e $\sigma = -4/7$, per il Processo di Dirichlet $\alpha = 2.25$, per Pitman-Yor $\sigma = 0.315$ e $\alpha = 0.475$, per Gneden $\gamma = 0.525$ e infine per Stirling-Gamma $\alpha = 20$ e $\beta = 2$. Ricorriamo ancora una volta a tutte le impostazioni elencate durante la trattazione della prima rete per l'implementazione dell'Algoritmo 1. In primo luogo analizziamo la distribuzione a posteriore del numero di gruppi per le diverse distribuzioni a priori, raffigurate nei grafici in Figura 3.35, Figura 3.36, Figura 3.37, Figura 3.38 e Figura 3.39. I cambiamenti apportati alla matrice a blocchi Θ hanno portato ad un diminuzione della variabilità per tutte le distribuzioni rispetto all'ultima rete analizzata. Osserviamo inoltre che tutte le distribuzioni hanno la moda sul valore 10, effettivo numero di gruppi. Continuiamo calcolando la partizione ottima \mathbf{z}^* per ogni distribuzione a priori attraverso l'algoritmo SALSO con funzione di perdita di Binder di parametro $a = 1$; tutte coincidono con la vera partizione. Infine confrontiamo le performance dei vari modelli ottenuti con il criterio WAIC, per capire quale distribuzione a priori produce il modello migliore in questo contesto. I valori del WAIC sono riportati nella Tabella 3.7. Notiamo che i valori sono tutti vicini tra di loro, in particolare tra il valore per la distribuzione a priori di Gneden, che è il più piccolo, e quello di Pitman-Yor, il più grande, intercorre una differenza si soli 0.53, ancora una volta questa differenza potrebbe essere semplicemente dovuta al campionamento Monte Carlo.

Osserviamo infine come cambiano i risultati per una combinazione degli iperparametri delle distribuzioni a priori tale che il numero atteso di gruppi sia 15. Per ottenere ciò impostiamo $\bar{H} = 26$ e $\sigma = -7/26$ per la distribuzione a priori

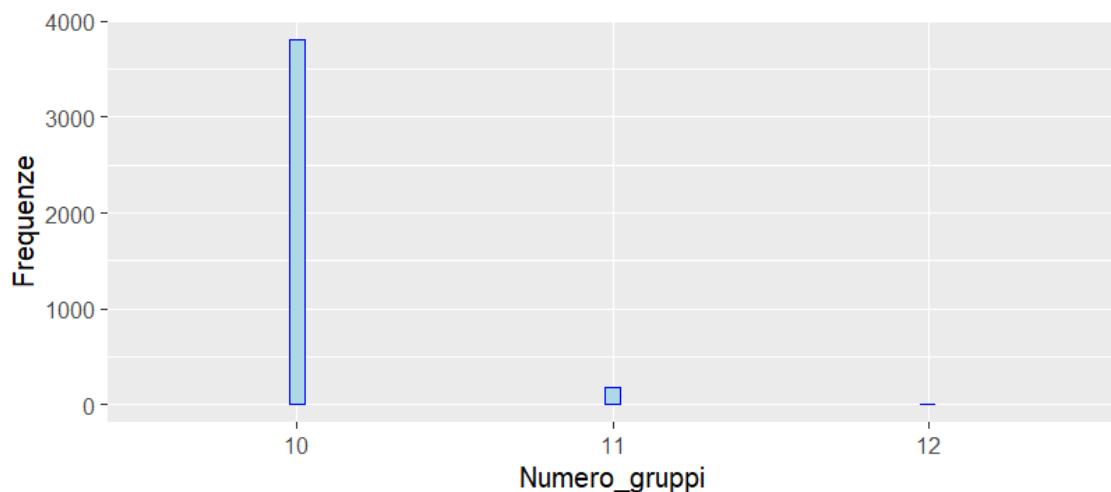


Figura 3.36: Distribuzione a posteriore del numero di gruppi, utilizzando la distribuzione a priori Processo di Dirichlet di parametro $\alpha = 2.25$ sulla rete costituita da dieci gruppi.

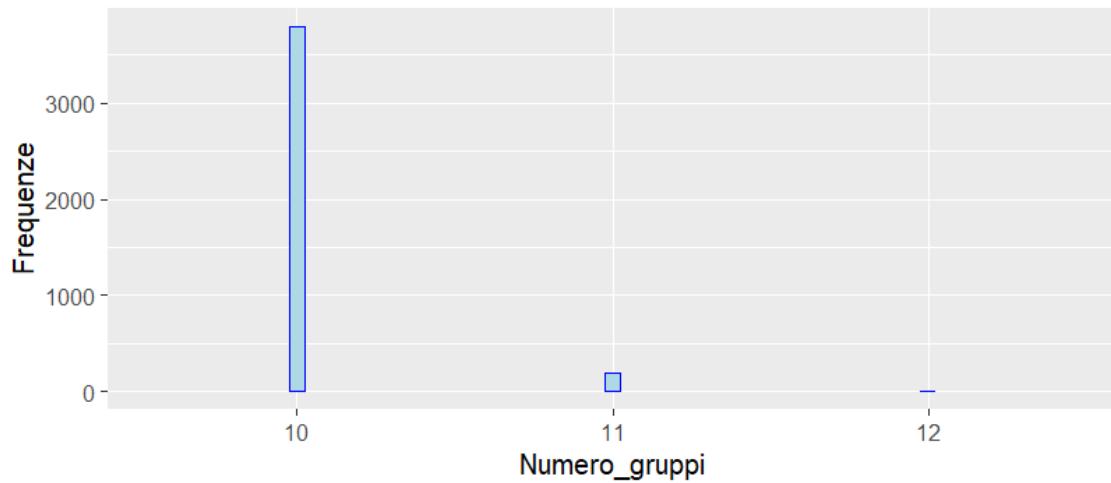


Figura 3.37: Distribuzione a posteriore del numero di gruppi, utilizzando la distribuzione a priori Pitman-Yor di parametri $\sigma = 0.315$ e $\alpha = 0.475$ sulla rete costituita da dieci gruppi.

WAIC	
Dirichlet-Multinomiae	11088.37
Processo di Dirichlet	11088.57
Pitman-Yor	11088.89
Gnedin	11088.36
Stirling-Gamma	11088.56

Tabella 3.7: WAIC calcolato sui campioni \mathbf{z} ottenuti dalla rete costituita da dieci gruppi, impiegando le diverse distribuzione a priori di tipo Gibbs con iperparametri che restituiscono 10 come valore atteso del numero di gruppi

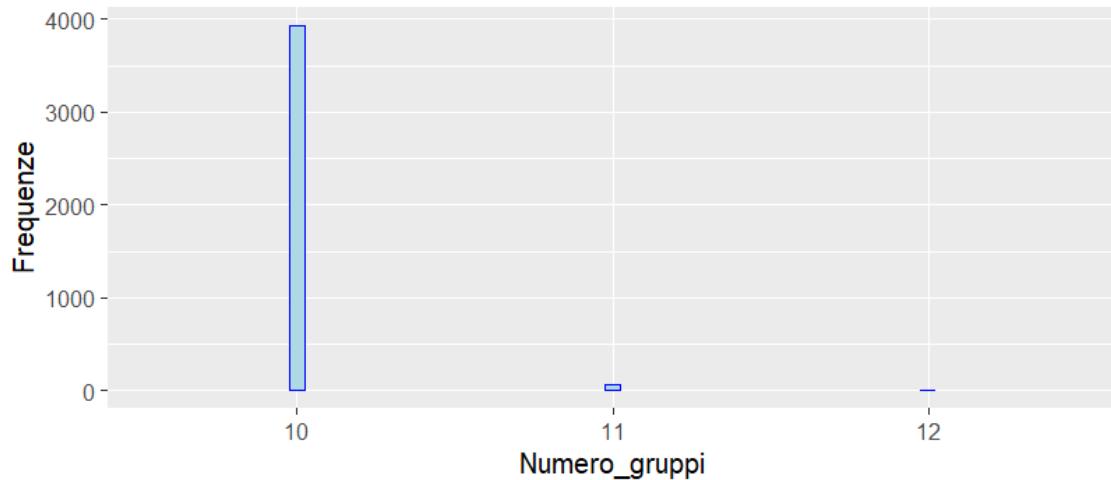


Figura 3.38: Distribuzione a posteriore del numero di gruppi, utilizzando la distribuzione a priori di Gnedin di parametro $\gamma = 0.525$ sulla rete costituita da dieci gruppi.

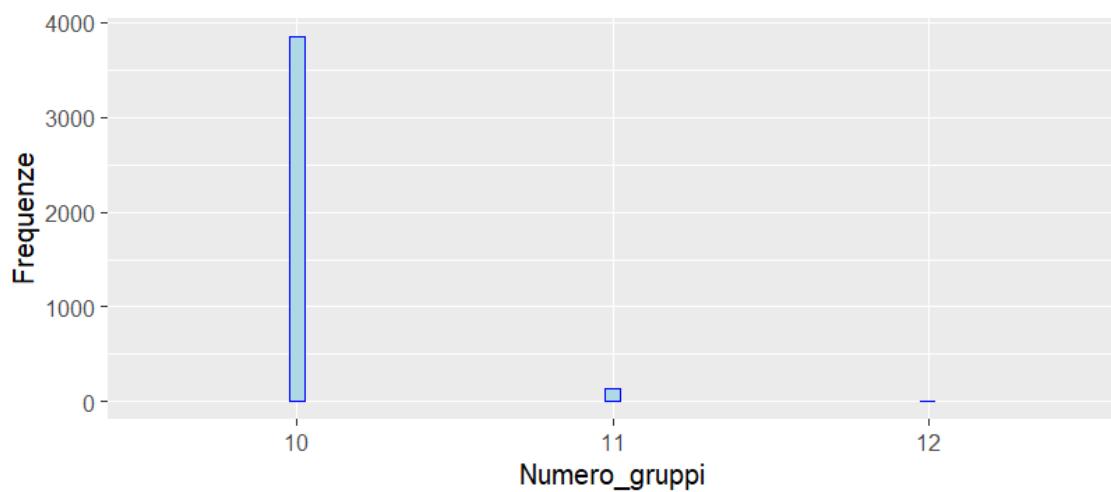


Figura 3.39: Distribuzione a posteriore del numero di gruppi, utilizzando la distribuzione a priori Stirling-Gamma di parametri $\alpha = 20$ e $\beta = 2$ sulla rete costituita da dieci gruppi.

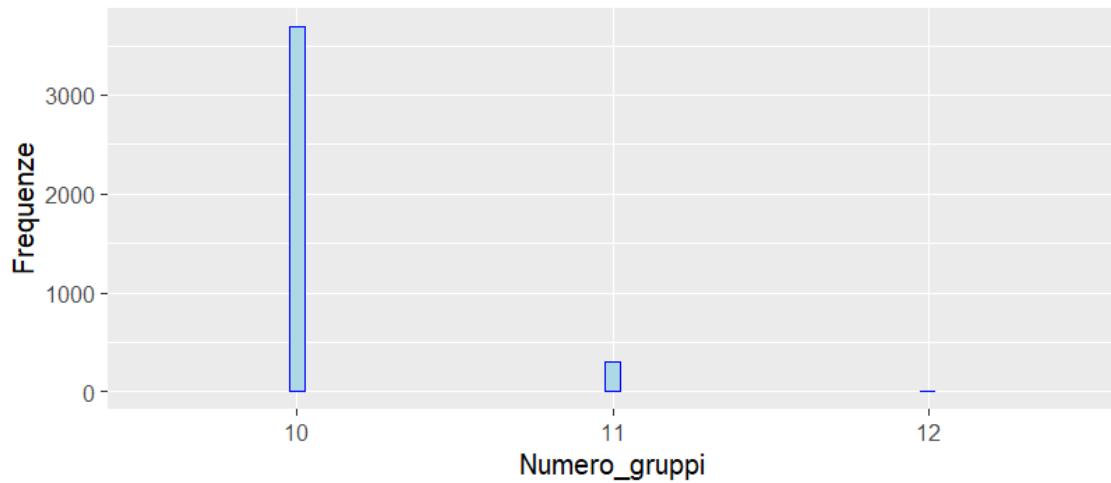


Figura 3.40: Distribuzione a posteriore del numero di gruppi, utilizzando la distribuzione a priori Dirichlet-Multinomiale di parametri $\bar{H} = 26$ e $\sigma = -7/26$ sulla rete costituita da dieci gruppi.

	WAIC
Dirichlet-Multinomilae	11089.89
Processo di Dirichlet	11088.91
Pitman-Yor	11088.86
Gnedin	11088.45
Stirling-Gamma	11088.82

Tabella 3.8: WAIC calcolato sui campioni \mathbf{z} ottenuti dalla rete costituita da dieci gruppi, impiegando le diverse distribuzioni a priori di tipo Gibbs con iperparametri che restituiscono 15 come valore atteso del numero di gruppi

Dirichlet-Multinomiale, $\alpha = 4$ per il Processo di Dirichlet, $\sigma = 0.415$ e $\alpha = 0.525$ per Pitman-Yor, $\gamma = 0.44$ per Gnedin e infine $\alpha = 30$ e $\beta = 2$ per Stirling-Gamma. Partiamo dall’analisi delle distribuzioni a posteriore del numero di gruppi, rappresentate nei grafici in Figura 3.40, Figura 3.41, Figura 3.42, Figura 3.43 e Figura 3.44. Non notiamo evidenti aumenti di variabilità, come avevamo di solito osservato aumentando il valore atteso del numero di gruppi rispetto al valore reale. Per quanto riguarda la partizione ottima \mathbf{z}^* , ancora una volta questa, calcolata tramite l’algoritmo SALSO con funzione di perdita di Binder di parametro $a = 1$, coincide con la partizione vera per tutte le distribuzioni a priori impiegate. Confrontiamo, infine, la performance dei modelli, i valori dell’indice WAIC per ogni distribuzione a priori si trovano nella Tabella 3.8. Il modello prodotto dalla distribuzione a priori di Gnedin si riconferma quello con valore inferiore di WAIC, mentre quello originato dalla distribuzione a



Figura 3.41: Distribuzione a posteriore del numero di gruppi, utilizzando la distribuzione a priori Processo di Dirichlet di parametro $\alpha = 4$ sulla rete costituita da dieci gruppi.

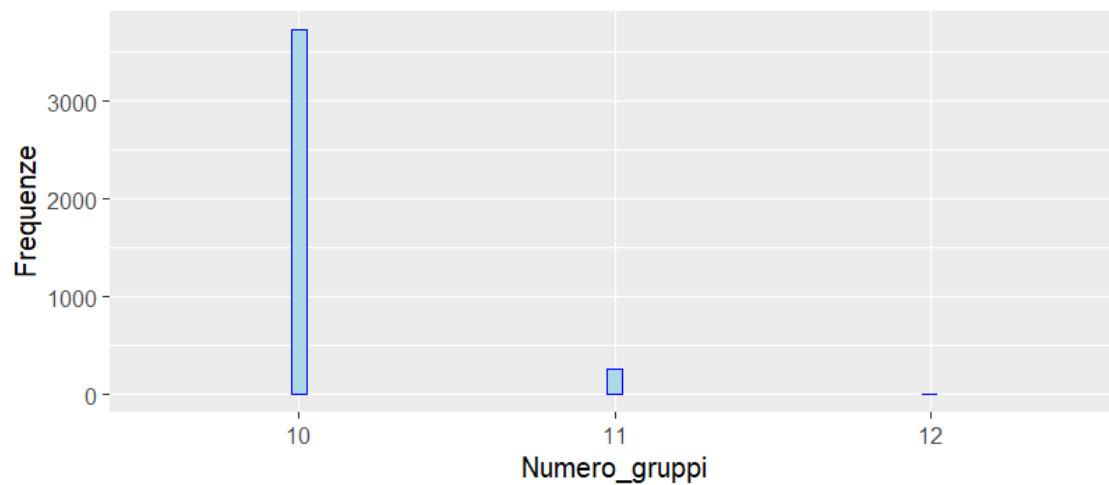


Figura 3.42: Distribuzione a posteriore del numero di gruppi, utilizzando la distribuzione a priori Pitman-Yor di parametri $\sigma = 0.415$ e $\alpha = 0.525$ sulla rete costituita da dieci gruppi.

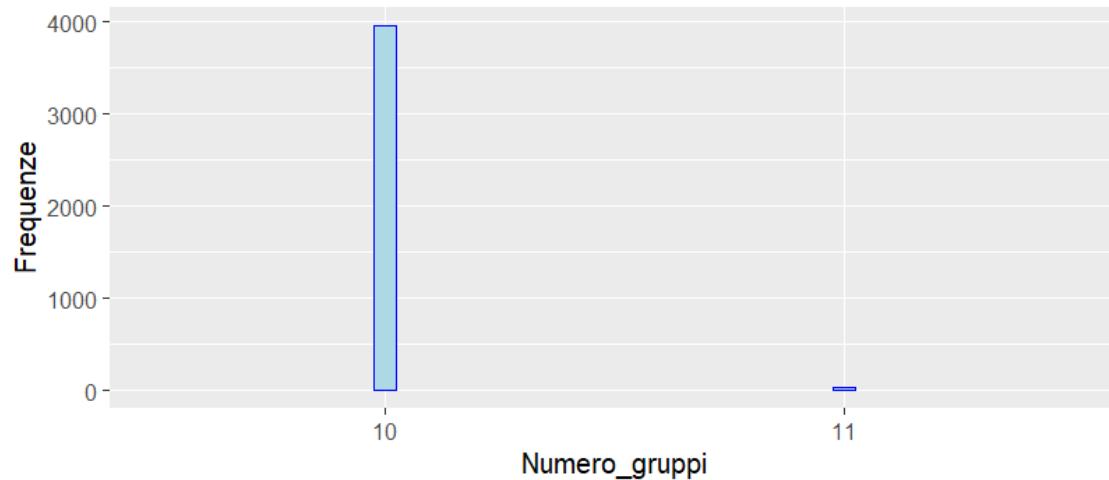


Figura 3.43: Distribuzione a posteriori del numero di gruppi, utilizzando la distribuzione a priori di Gnedin di parametro $\gamma = 0.44$ sulla rete costituita da dieci gruppi.

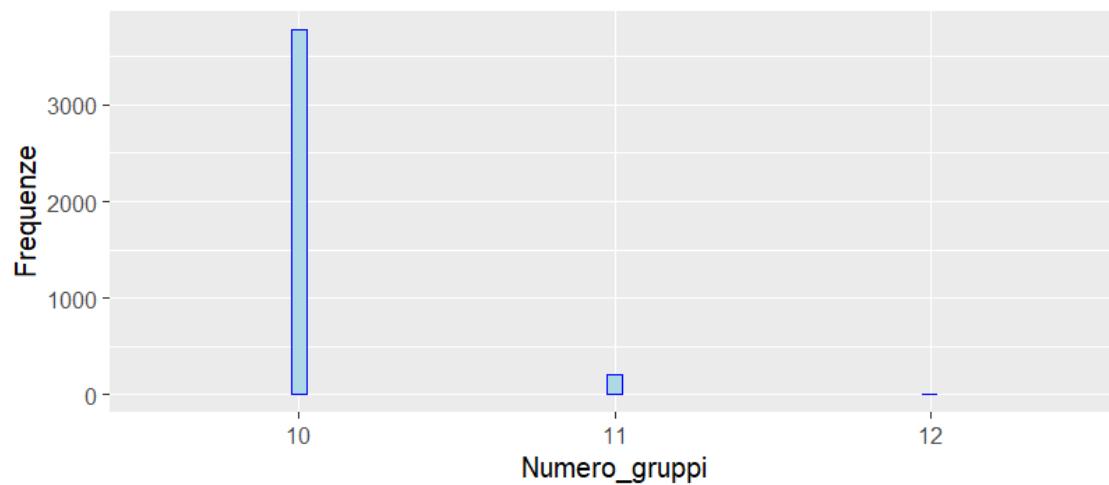


Figura 3.44: Distribuzione a posteriori del numero di gruppi, utilizzando la distribuzione a priori Stirling-Gamma di parametri $\alpha = 30$ e $\beta = 2$ sulla rete costituita da dieci gruppi.

priori Dirichlet-Multinomiale assume il valore più alto, discostandosi di circa un punto dagli altri.

In conclusione abbiamo riscontrato il buon funzionamento dei modelli per blocchi stocastici estesi applicati a questi semplici esempi simulati. In particolare abbiamo potuto osservare come le distribuzioni a priori di Gnedin e di Stirling-Gamma siano quelle meno influenzate dal cambio degli iperparametri, questo conferma l'effetto positivo dell'impiego di una *iperprior*, rispettivamente sul numero di gruppi \bar{H} e sul parametro α del processo di Dirichlet. Inoltre abbiamo potuto constatare come l'impiego dell'algoritmo SALSO permetta, grazie all'utilizzo della versione generalizzata della funzione di perdita di Binder, di controllare il numero di gruppi contenuti nella partizione ottima \mathbf{z}^* , a differenza dell'algoritmo implementato da Wade & Ghahramani (2018).

3.2 Applicazione su rete che simula le caratteristiche della rete Infinito

Andiamo adesso a valutare la performance dei modelli per blocchi stocastici estesi in un contesto che mimi la nostra applicazione reale. Consideriamo dunque una rete composta da $W = 80$ nodi, campionata da un modello per blocchi stocastici con $H_0 = 5$ gruppi, dove i primi 25 nodi appartengono al primo gruppo, i seguenti 20 al secondo, i successivi 15 al terzo, i 15 dopo al quarto e gli ultimi 5 al quinto. La probabilità dei blocchi Θ è pari 0.75 per due nodi che appartengono allo stesso gruppo e 0.25 altrimenti:

$$\Theta = \begin{bmatrix} 0.75 & 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.75 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.75 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.75 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 & 0.75 \end{bmatrix}$$

La matrice di adiacenza simulata con queste caratteristiche è rappresentata in Figura 3.45, nella quale i colori laterali corrispondono alla vera partizione \mathbf{z}_0 , le celle nere si riferiscono ai legami, mentre quelle bianche ai non legami. Questa rete assomiglia alla struttura a blocchi della rete infinito, nella quale ci aspettiamo strutture comunitarie tra gli affiliati di ciascuna *locale*, strutture composte da nuclei periferici tra affiliati e i loro boss, risultanti dalle strategie di copertura messe in atto dall'organizzazione.

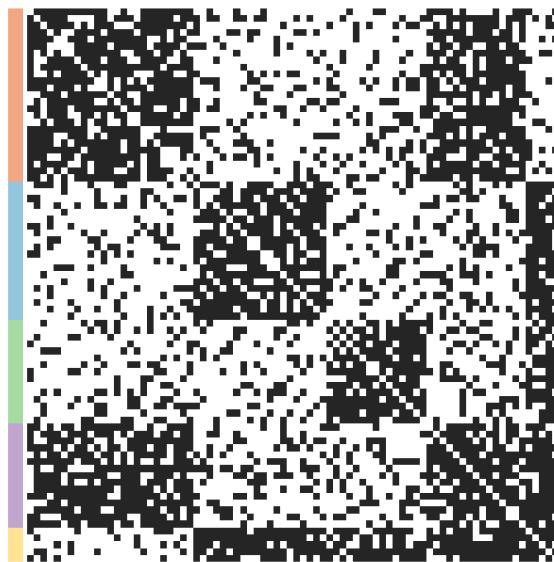


Figura 3.45: Matrice di adiacenza simulata

Andiamo adesso ad implementare i modelli per blocchi stocastici estesi , sia senza che con gli attributi dei nodi, che in questo caso coincidono con la vera partizione \mathbf{z}_0 . Tale scelta è utile per valutare fino a che punto le distribuzioni a priori di tipo Gibbs riescono a sfruttare efficacemente attributi informativi per i nodi. All'interno della classe di tipo Gibbs iniziamo valutando le cinque distribuzioni a priori non supervisionate per $p(\mathbf{z})$: Dirichlet-Multinomiale, Processo di Dirichlet, Pitman-Yor, Gneden e Stirling-Gamma. Impostiamo gli iperparametri in modo tale che il numero atteso di gruppi a priori sotto le diverse distribuzioni a priori sia circa $H_0 = 10$, mentre inizializziamo l'Algoritmo 1 con ogni nodo in un gruppo diverso, in questo modo possiamo controllare la robustezza dei risultati per questa specificazione dei parametri e per l'inizializzazione del campionamento Gibbs. Nello specifico, poniamo $\bar{H} = 50$ e $\sigma = -3.5/50$ per la distribuzione a priori Dirichlet-Multinomiale, $\alpha = 3$ per il Processo di Dirichlet, $\sigma = 0.6$ e $\alpha = -0.3$ per Pitman-Yor, $\gamma = 0.45$ per Gneden e $\alpha = 20$ e $\beta = 2$ per Stirling-Gamma. Nell'implementare tali modelli consideriamo il contesto standard uniforme $a = b = 1$ per la *prior* sulle probabilità dei blocchi. Per l'Algoritmo 1 otteniamo 40000 campioni per \mathbf{z} , dopo un *burn-in* conservativo di 10000.

Passiamo ora all'analisi dei risultati. Iniziamo calcolando la partizione ottima, per fare ciò siamo ricorsi ancora una volta all'algoritmo SALSO con funzione di perdita di Binder di parametro $a = 1$. Tutte le distribuzioni a priori impiegate hanno individuato la stessa partizione ottima caratterizzata da 5 gruppi, in cui i nodi appartenenti ai gruppi 2, 3 e 5 sono stati tutti correttamente assegnati,

	WAIC	$H_{0.25}$	$H_{0.5}$	$H_{0.75}$	$binder(\hat{\mathbf{z}}, \mathbf{z}_b)$
Dirichlet-Multinomiae	3543.6	5	6	6	0.085
Processo di Dirichlet	3543.3	5	6	6	0.083
Pitman-Yor	3543.7	5	6	6	0.08
Gnedin	3536.4	5	5	5	0.075
Stirling-Gamma	3541.39	5	6	6	0.0832

Tabella 3.9: Performance dei ESBM per la rete che mima la *rete Infinito*, con $H_0 = 5$ numero reale di gruppi, nel caso in cui non si includano gli attributi dei nodi

mentre un nodo del gruppo 1 è stato assegnato al quarto e due di quest'ultimo sono stati attribuiti al primo. Continuiamo confrontando le performance delle diverse distribuzioni a priori, attraverso il criterio d'informazione WAIC, i quartili della distribuzione a posteriori del numero di gruppi H e la distanza di Binder, $binder(\hat{\mathbf{z}}, \mathbf{z}_b)$ tra la partizione stimata $\hat{\mathbf{z}}$ e il limite affidabile al 95% \mathbf{z}_b . I risultati di queste tre misure per ogni distribuzione a priori impiegata per il calcolo dei campioni \mathbf{z} , sono illustrati nella Tabella 3.9. Tra tutte le distribuzioni a priori non supervisionate di tipo Gibbs considerate per $p(\mathbf{z})$, il processo di Gnedin utilizzato come distribuzione a priori su questi modelli produce i risultati migliori in termini di WAIC, presenta infatti il valore inferiore. Inoltre, fornisce un apprendimento del numero di gruppi più accurato, in quanto restituisce un range interquartile più stretto attorno al vero valore $H_0 = 5$ per la distribuzione a posteriori del numero di gruppi. Infine mostra un *credible ball* più stretto attorno alla partizione ottima $\hat{\mathbf{z}}$.

Concludiamo l'analisi di questa rete implementando nuovamente l'Algoritmo 1, questa volta aggiungendo l'informazione che abbiamo sui nodi degli attributi, impostiamo perciò i parametri necessari come segue: $\alpha_1 = \dots = \alpha_C = 1$. Per quanto riguarda la partizione ottima, questa coincide per tutte le distribuzioni a priori impiegate con quella vera, questo dimostra un miglioramento dei nostri modelli come conseguenza dell'aggiunta degli attributi dei nodi. Questo miglioramento è confermato anche dal criterio d'informazione WAIC, i quartili della distribuzione a posteriori del numero di gruppi H e la distanza di Binder, $binder(\hat{\mathbf{z}}, \mathbf{z}_b)$ tra la partizione stimata $\hat{\mathbf{z}}$ e il limite affidabile al 95% \mathbf{z}_b , questi valori, rappresentati nella Tabella 3.10, infatti, sono preferibili a quelli in Tabella 3.9. Notiamo come dopo l'inclusione degli attributi dei nodi, è sempre la distribuzione a priori di Gnedin quella che mostra le performance migliori. Inoltre possiamo constatare come l'utilizzo della *iperprior* Stirling-Gamma per il parametro α del Processo di Dirichlet produca risultati migliori per gli indici considerati, rispetto a tenere fisso il parametro, sia

	WAIC	$H_{0.25}$	$H_{0.5}$	$H_{0.75}$	$binder(\hat{\mathbf{z}}, \mathbf{z}_b)$
Dirichlet-Multinomiae	3520.8	5	6	6	0.035
Processo di Dirichlet	3519.4	5	5	6	0.035
Pitman-Yor	3517.2	5	5	6	0.31
Gnedin	3515.4	5	5	5	0.025
Stirling-Gamma	3516.4	5	5	5	0.031

Tabella 3.10: Performance dei ESBM per la rete che mima la *rete Infinito*, con $H_0 = 5$ numero reale di gruppi, nel caso in cui si includano gli attributi dei nodi

prima che dopo l'inclusione degli attributi dei nodi. Per questo motivo illustriamo la matrice di similarità a posteriori prodotta dai modelli per blocchi stocastici estesi per queste due distribuzioni a priori, sia nel caso in cui non si includano gli attributi dei nodi che in quello in cui invece vengono utilizzati, in Figura 3.46. In questa rappresentazione la prima riga si riferisce alla distribuzione a priori di Gnedin, mentre la seconda alla Stirling-Gamma, le colonne invece fanno riferimento alla presenza (destra) o assenza (sinistra) degli attributi dei nodi nell'implementazione del modello. Il colore delle celle varia dal bianco al nero come la probabilità dei gruppi stimata per due nodi in contemporanea va da 0 a 1, i colori laterali corrispondono alla stima della partizione ottima $\hat{\mathbf{z}}$. Possiamo osservare come per entrambe le due distribuzioni a priori impiegate per $p(\mathbf{z})$ la matrice di similarità a posteriori migliori dopo l'aggiunta dell'informazione sui nodi, inoltre grazie a questa si riducono anche le differenze tra le due *prior*.

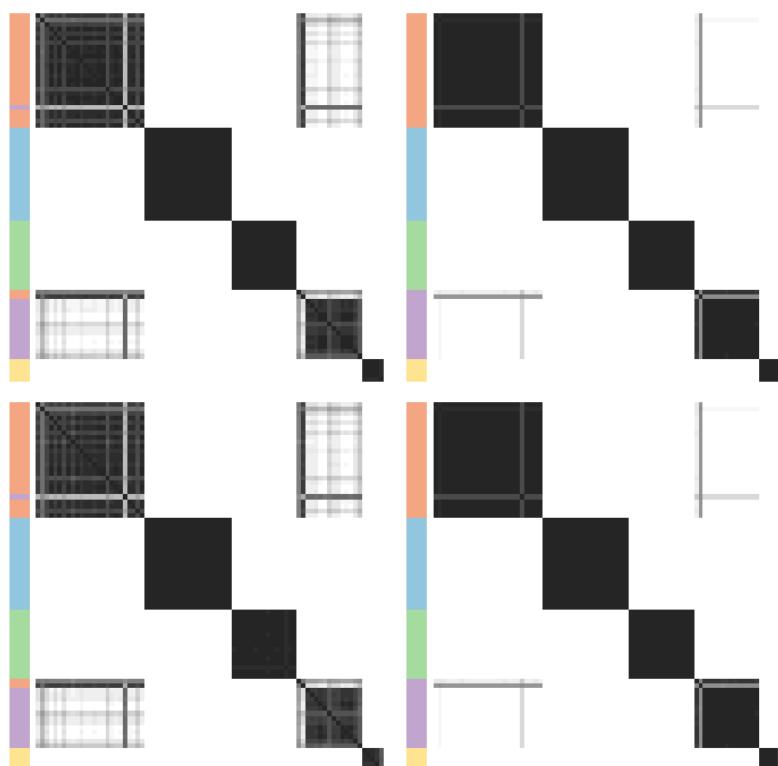


Figura 3.46: Matrice di similarità a posteriori dei ESBM per le distribuzioni a priori di Gnedin e Stirling Gamma con e senza gli attributi dei nodi

Capitolo 4

Applicazione dei ESBM alla *rete Infinito*

Concludiamo questo elaborato andando ad applicare i modelli per blocchi stocastici estesi sulla *rete Infinito*, per comprendere la struttura interna de *La Lombardia*. In primo luogo andiamo ad analizzare i risultati ottenuti senza inserire nei modelli gli attributi specifici dei nodi, che indicano la *locale* di appartenenza e il ruolo all'interno dell'organizzazione per ciascun affiliato secondo gli atti giudiziari. In secondo luogo andiamo a vedere come migliorano i nostri modelli includendo anche queste informazioni. Infine andiamo a descrivere la composizione de *La Lombardia* utilizzando il modello che ha mostrato il comportamento migliore.

4.1 Performance dei ESBM applicati alla *rete Infinito* senza l'inclusione degli attributi dei nodi

Per comprendere in profondità la struttura interna sottostante *La Lombardia*, attuiamo un'analisi della *rete Infinito* utilizzando la classe dei modelli per blocchi stocastici estesi, inizialmente senza considerare le informazioni aggiuntive che abbiamo sui nodi. L'obiettivo di questa prima analisi consiste nell'identificare il modello che si comporta meglio sui dati che abbiamo a disposizione, comparando le performance delle distribuzioni a priori non supervisionate per \mathbf{z} con iperparametri che inducono il valore atteso a posteriori del numero di gruppi ad essere 20. Questo valore è quattro volte superiore al numero di locali presenti nella rete, perciò sembra ragionevolmente conservativo. Poniamo dunque $\bar{H} = 50$ e $\sigma = -12/50$ per la distribuzione a priori Dirichlet-Multinomiale, $\alpha = 8$ per il Processo di Dirichlet, $\sigma = 0.725$ e $\alpha = -0.350$ per Pitman-Yor, $\gamma = 0.2$ per Gnedenko e $\alpha = 40$, $\beta = 2$ per Stirling-Gamma. L'inferenza a posteriori si basa, come nell'ultimo esempio simulato, su 40000 campioni MCMC prodotti dall'Algoritmo 1, dove inizialmente ogni nodo apparteneva ad un gruppo

	WAIC	$H_{0.25}$	$H_{0.5}$	$H_{0.75}$	$binder(\hat{\mathbf{z}}, \mathbf{z}_b)$
Dirichlet-Multinomiae	1235.5	14	15	15	0.016
Processo di Dirichlet	1255.1	14	15	15	0.019
Pitman-Yor	1271.7	13	14	15	0.008
Gnedin	1217.8	15	15	15	0.017
Stirling-Gamma	1264.8	14	14	15	0.027

Tabella 4.1: Performance dei ESBM per sulla *rete Infinito*, nel caso in cui non si includano gli attributi dei nodi

differente e considerando il contesto standard uniforme per la distribuzione a priori della probabilità dei blocchi: $a = b = 1$, dopo un periodo di *burn-in* di 10000. Una volta ottenuti tutti i campioni MCMC, andiamo a calcolare la partizione ottima grazie all’algoritmo SALSO con funzione di perdita di Binder di parametro $a = 1$. Questa la utilizzeremo per calcolare la distanza di Binder, $binder(\hat{\mathbf{z}}, \mathbf{z}_b)$, tra la partizione stimata $\hat{\mathbf{z}}$ e il limite affidabile al 95% \mathbf{z}_b . Questo valore, insieme all’indice WAIC e ai quartili della distribuzione a posteriori del numero di gruppi, calcolati per ogni distribuzione a priori, sono riportati nella Tabella 4.1. Possiamo notare che il modello che mostra il valore di WAIC minore è quello che ricorre alla distribuzione a priori di Gnedin, come nell’esempio simulato analizzato nel capitolo precedente.

4.2 Performance dei ESBM applicati alla *rete Infinito* includendo gli attributi dei nodi

Andiamo adesso ad implementare i modelli per blocchi stocastici estesi includendo le informazioni che abbiamo sui nodi: *locale* di appartenenza e ruolo all’interno dell’organizzazione criminale, secondo gli atti giudiziari. Per fare ciò impostiamo i parametri delle distribuzioni a priori di tipo Gibbs come nel caso precedente in cui non abbiamo utilizzato gli attributi dei nodi, inoltre poniamo i parametri della distribuzione di \mathbf{X}_h , $\alpha_1, \dots, \alpha_C$, tutti pari a 1. Calcoliamo la partizione ottima nuovamente su 40000 campioni MCMC prodotti dall’Algoritmo 1, dopo un periodo di *burn-in* di 10000, ricorrendo all’algoritmo SALSO con funzione di perdita di Binder di parametro $a = 1$. Vediamo come sono cambiati i valori di WAIC, quartili della distribuzione a posteriori del numero di gruppi e la distanza di Binder tra la partizione stimata e il limite affidabile al 95%, nella Tabella 4.2. Possiamo notare che la distribuzione a priori di Gnedin riporta le performance migliori nella *rete Infinito* rispetto alle altre *prior* di tipo Gibbs impiegate. Questo fornisce un supporto

	WAIC	$H_{0.25}$	$H_{0.5}$	$H_{0.75}$	$binder(\hat{\mathbf{z}}, \mathbf{z}_b)$
Dirichlet-Multinomiae	1221.3	14	14	15	0.007
Processo di Dirichlet	1212.7	15	15	15	0.009
Pitman-Yor	1264.6	14	14	15	0.014
Gnedin	1198.4	15	15	16	0.013
Stirling-Gamma	1202.7	15	15	15	0.009

Tabella 4.2: Performance dei ESBM per sulla *rete Infinito*, nel caso in cui si includano gli attributi dei nodi

quantitativo a quanto detto in precedenza riguardo la distribuzione a priori di Gnedin, ovvero che questa sia una *prior* realistica per capire da quali gruppi sia strutturata un’organizzazione criminale. Inoltre, confrontando la Tabella 4.1 con la Tabella 4.2, osserviamo che le distribuzioni a priori supervisionate grazie all’informazione aggiuntiva sul ruolo e sulla *locale* d’affiliazione portano ad una riduzione del WAIC e, quasi in tutti i casi, ad un abbassamento dell’incertezza a posteriori, ciò significa che gli attributi inseriti nei modelli contengono informazioni sui moduli che costituiscono la ’Ndrangheta.

4.3 Analisi della struttura interna de *La Lombardia* attraverso il ESBM con distribuzione a priori di Gnedin supervisionata

Concludiamo l’analisi della *rete Infinito* andando ad analizzare la partizione ottima \mathbf{z}^* identificata dall’algoritmo SALSO per la distribuzione a priori di Gnedin, così da tessere un disegno della struttura interna de *La Lombardia*.

Ricordando le principali teorie sul crimine organizzato, la nostra ipotesi è che la ’Ndrangheta mostri strutture a blocchi complesse, in cui le comunità di affiliati appartenenti a ciascuna *locale* sono combinate con strutture a blocchi connesse ad un livello di coordinamento superiore: quello tra i boss. A differenza dei classici algoritmi di rilevamento di comunità, i modelli per blocchi stocastici estesi tengono particolarmente conto di queste architetture, fornendo perciò evidenza empirica senza precedenti a supporto di tali teorie forensi. Andiamo a visualizzare quanto detto grazie alla Figura 4.1 e alla Figura 4.2. Queste rappresentazioni grafiche si basano sulla stima puntuale della partizione ottima \mathbf{z}^* sotto la distribuzione a priori di Gnedin supervisionata, questa sarà la *prior* che verrà considerata da questo momento in poi per tutta la l’analisi della *rete Infinito*, per via della sua performance superiore,

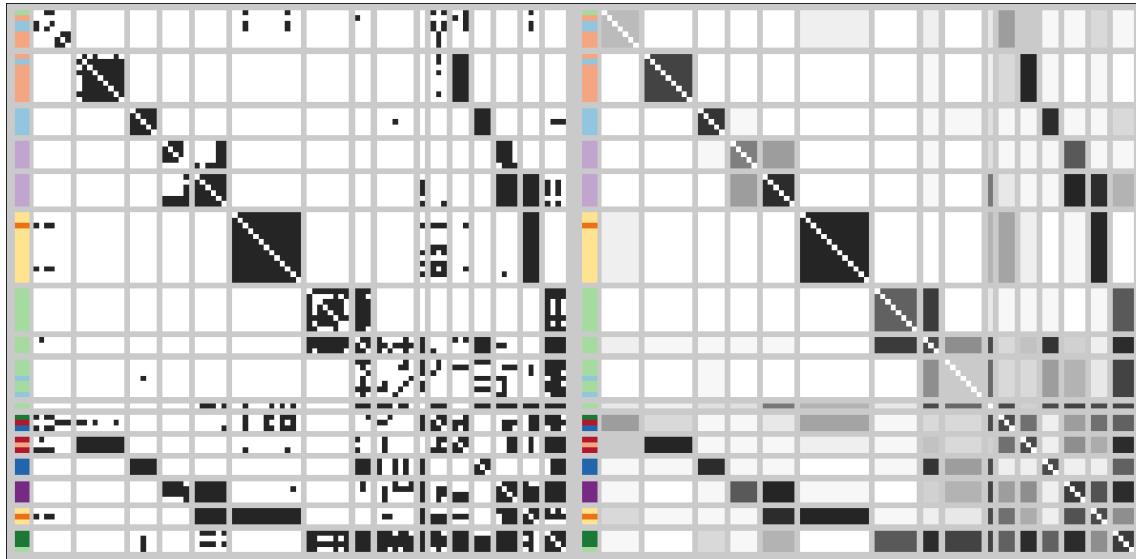


Figura 4.1: Matrice di adiacenza e stima della matrice delle probabilità dei legami

evidenziata nella Tabella 4.2 e l’incertezza a posteriori attorno alla stima puntuale \mathbf{z}^* relativamente bassa. Le rappresentazione grafiche in Figura 4.1 raffigurano la matrice di adiacenza (sinistra) e la stima della matrice delle probabilità dei legami (destra) della *rete Infinito*, con i nodi ri-ordinati e partizionati in blocchi secondo la struttura di clustering stimata grazie ai modelli per blocchi stocastici estesi con distribuzione a priori di Gnedin. I colori laterali corrispondono alle diverse *locali*, mentre i toni più scuri o chiari indicano rispettivamente i boss e gli affiliati. Mentre in Figura 4.2 è rappresentata la rete dei gruppi stimati nella *rete Infinito*, dove ogni nodo rappresenta un gruppo e i legami sono pesati dalle probabilità dei blocchi stimate. Le dimensioni dei nodi sono proporzionali alle cardinalità dei gruppi, mentre i grafici a torta rappresentano la loro composizione che tiene conto della *locale* di appartenenza e il ruolo ricoperto all’interno dell’organizzazione, i colori corrispondono anche in questo caso alle diverse *locali*, mentre la tonalità, scura o chiara, indica rispettivamente i boss o gli affiliati. Per fornire informazioni più dirette, la composizione relativa al ruolo nei grafici a torta di dimensione più piccola è ripesata per tenere conto del fatto che i boss sono meno frequenti nella rete rispetto agli affiliati. Le posizioni dei nodi sono ottenute attraverso un collocamento forzato e orientato (Fruchterman & Reingold, 1991), per riflettere la forza delle connessioni.

Per confermare ulteriormente le ipotesi forensi, calcoliamo la differenza tra il WAIC nel caso non supervisionato e supervisionato per la distribuzione a priori di Gnedin. L’attributo di ciascun sospettato, \mathbf{X} , definisce la struttura ipotizzata, in particolare la classe di ciascun affiliato corrisponde alla *locale* associata, mentre

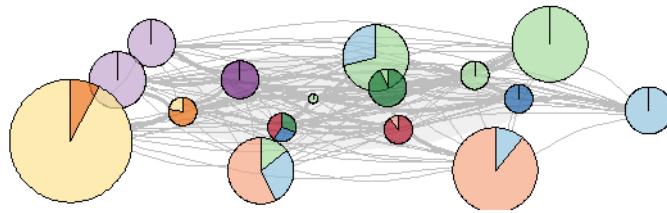


Figura 4.2: Rappresentazione della rete dei gruppi stimati nella *rete Infinito*

tutti i boss hanno in comune un'etichetta che indica che tali membri hanno un ruolo di leadership all'interno dell'organizzazione. Inoltre un sottoinsieme di affiliati appartenenti alla *locale D*, raffigurata con il colore viola, risulta, dagli atti giudiziari, ricoprire un ruolo periferico e per questo sono indicati da un'etichetta differente. La differenza tra i due WAIC è di circa 19, questo fornisce una forte evidenza a favore della nostra ipotesi, quando comparato con la soglia suggerita per tali criteri d'informazione.

La distribuzione a priori supervisionata di Gnedin è la migliore per apprendere le strutture a blocchi che caratterizzano la *rete Infinito*. Secondo la Figura 4.1 e la Figura 4.2, i moduli che possiamo osservare suggeriscono una struttura caratterizzata da partizioni annidate, principalmente definite da due macro blocchi di affiliati e boss, i quali sono ulteriormente partizionati in sotto gruppi coerenti principalmente con la *locale*, questi a loro volta mostrano strutture assortative deboli nell'architettura di coordinamento di livello superiore tra i boss e le differenti *locali*. Come conferma, andiamo a rappresentare il grafico a dispersione della *betweenness* media e della transitività locale per ogni gruppo stimato con la distribuzioni a priori di Gnedin supervisionata in Figura 4.3. Le dimensioni sono proporzionali alle cardinalità dei gruppi, mentre il colore di ciascun punto è posto uguale a quello che occupa la porzione più grande nel grafico a torta associato, rappresentato nella Figura 4.2. I cerchi e i quadrati rappresentano rispettivamente i gruppi che si riferiscono principalmente agli affiliati e ai boss. La Figura 4.3 mostra come i gruppi composti da affiliati siano tipicamente caratterizzati da una transitività locale alta e da una *betweenness* bassa, mentre i gruppi costituiti dai boss mostrano il comportamento opposto. Questa

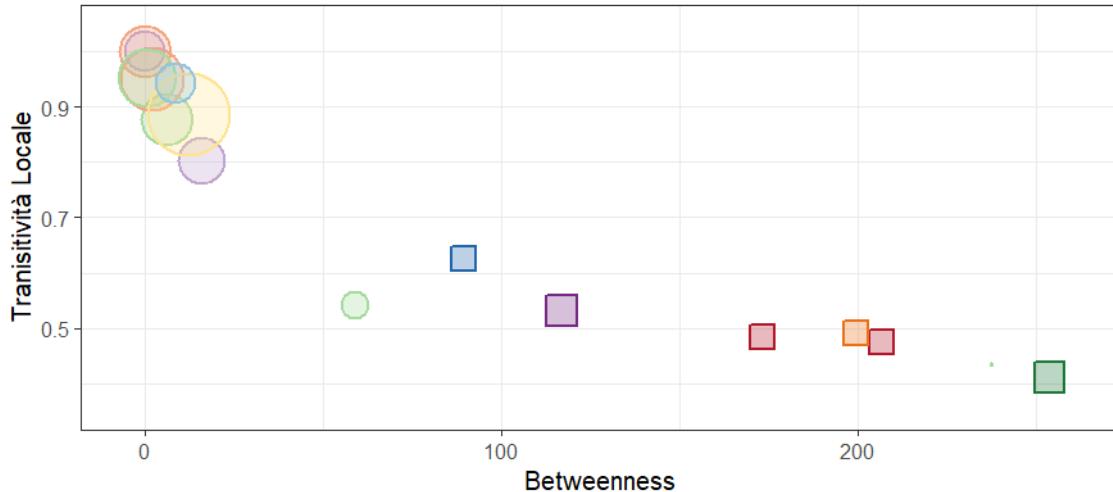


Figura 4.3: Grafico a dispersione della *betweenness* media e della transitività locale per ogni gruppo stimato grazie alla distribuzione a priori di Gneden supervisionata

scoperta è fondamentale in quanto fornisce una nuova evidenza empirica sul tentativo dei boss della 'Ndrangheta di affrontare il compromesso tra efficienza e sicurezza attraverso la creazione di gruppi centrali di piccola dimensione e sparsi, caratterizzati da una *betweenness* alta che favorisce il flusso d'informazione attraverso i gruppi più grandi e densi costituiti dagli affiliati, i quali garantiscono l'efficienza. Accanto a queste architetture ricorrenti, la flessibilità dei modelli per blocchi stocastici estesi riesce anche a tenere conto di altre deviazioni locali informative. Ad esempio, il primo gruppo rappresentato nella 4.1 comprende affiliati provenienti da diverse *locali*, i quali ricoprivano ruoli periferici secondo gli atti giudiziari. Similmente, gli schemi di connessione moderati tra i blocchi delle *locali* D e E, raffigurate rispettivamente con il colore viola e giallo, nella Figura 4.1 e nella Figura 4.2, sono consistenti con il fatto che la *locale* E è stata creata come ramificazione della *locale* D. La *locale* C, caratterizzata nei grafici attraverso il colore verde, ha una struttura a blocchi più complessa tra gli affiliati, a causa di una frammentazione in vari sotto gruppi che indicano posizioni di comando di livelli intermedio. Secondo gli atti giudiziari queste posizioni si riferiscono tipicamente a ruoli autoritari di supervisione durante le azioni criminali, che permettono di garantire un coordinamento tra *La Lombardia* e le famiglie che guidano la 'Ndrangheta in Calabria. Ruoli simili sono occupati anche da piccole frazioni di affiliati allocati in gruppi costituiti da boss. Tra questi affiliati vale la pena evidenziare il sospetto collocato in un gruppo composto da un solo nodo che occupa la posizione più centrale nella Figura 4.2. Nonostante questo non sia classificato come boss negli atti giudiziari, tale sospettato risulta essere un membro

anziano di alto rango all'interno dell'organizzazione, che ricopre ruoli fondamentali di mediazione tra tutte le *locali* e le famiglie a capo della 'Ndrangheta in Calabria. L'attuale posizione di tale affiliato nella struttura verticale de *La Lombardia* potrebbe essere quindi molto superiore rispetto a quanto è riportato correntemente.

Conclusione

In questo elaborato ci siamo concentrati sullo studio dei modelli per blocchi stocastici estesi, sviluppati per comprendere a fondo la complessa struttura sottostante le organizzazioni criminali, come la 'Ndrangheta. Le reti criminali forniscono un campo applicativo importante per gli sviluppi dei modelli per reti, ciò nonostante la loro considerazione nella letteratura statistica si è limitata a un numero ristretto di metodi, che offrivano una conoscenza parziale sulla struttura interna delle organizzazioni criminali. Per sopperire a questa mancanza, [Legramanti et al. \(2022\)](#) hanno appunto proposto i modelli per blocchi stocastici estesi, un'ampia classe di modelli pratici che unisce i già esistenti modelli per blocchi stocastici con le distribuzioni a priori di tipo Gibbs. Oltre a fornire uno schema metodologico, teorico e computazionale unico per i vari SBM, tale generalizzazione facilita la proposta di nuovi modelli, esplorando le opzioni alternative all'interno della classe di tipo Gibbs e permette la naturale inclusione di attributi attraverso la connessione con i PPM. Abbiamo infatti incluso in questo elaborato la distribuzione a priori di Stirling-Gamma appartenente alla classe di tipo Gibbs, non ancora presa in considerazione per questi modelli. Nel corso delle applicazioni descritte nei precedenti capitoli, abbiamo potuto osservare come questa nuova distribuzione abbia fornito risultati migliori rispetto a quelli ottenuti impiegando come distribuzione a priori il Processo di Dirichlet, in particolare quando combinata con gli attributi specifici dei nodi. La distribuzione a priori che fornisce la miglior partizione rimane il processo di Gnedin. Abbiamo inoltre proposto un nuovo algoritmo per il calcolo della partizione ottima, più flessibile di quello precedentemente utilizzato in questo contesto: l'algoritmo SALSO di [Dahl et al. \(2022\)](#). Questo permette, cambiando il parametro a , di ottenere partizioni diverse da poter confrontare con la nostra conoscenza sulla rete in esame. L'applicazione della rete *Infinito* ai modelli per blocchi stocastici estesi ha permesso di fornire evidenza empirica ai diverse teorie forensi riguardo il funzionamento interno di organizzazioni criminali complesse come la 'Ndrangheta.

Oltre all'introduzione di nuove distribuzioni a priori di tipo Gibbs, come la Stirling-Gamma, o l'impiego di algoritmo differenti per il calcolo della partizione ottima, come l'algoritmo SALSO, ulteriori estensioni sono possibili per questi modelli. Per esempio, nel caso si disponesse di informazioni aggiuntive sui nodi, oltre a quelle di tipo categoriale, queste potrebbero essere facilmente inserite modificando la distribuzione degli attributi, come descritto da Müller & Rosner (2011).

Bibliografia

- CALDERONI, F., BRUNETTO, D. & PICCARDI, C. (2017). Communities in criminal networks: A case study. *Soc. Netw.* **48**.
- DAHL, D. B., JOHNSON, D. J. & MÜLLER, P. (2022). Search algorithms and loss functions for bayesian clustering. *Journal of Computational and Graphical Statistics* **31**.
- DE BLASI, P., FAVARO, S., LIJOI, A., MENA, R. H., PRÜNSTER, I. & RUGGIERO, M. (2015). Are gibbs-type priors the most natural generalization of the dirichlet process? *IEEE Trans. Pattern Anal. Mach. Intell.* **37**.
- FERGUSON, T. S. (2011). A bayesian analysis of some nonparametric problems. *Ann. Statist* **1(2)**.
- FRUCHTERMAN, T. & REINGOLD, E. M. (1991). Graph drawing by force-directed placement. *Softw. Pract. Exp.* **21**.
- GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. & RUBIN, D. (2021). *Bayesian Data Analysis*. Chapman & Hall/CRC.
- GNEDIN, A. (2010). A species sampling model with finitely many types. *Electron. Commun. Probab.* **15**.
- KOLACZYK, E. D. & CSÁRDI, G. (2020). *Statistical Analysis of Network Data with R*. Springer.
- LEE, C. & WILKINSON, D. J. (2019). Review of stochastic block models and extensions for graph clustering. *EAppl. Netw. Sci.* **4**.
- LEGRAMANTI, S., RIGON, T., DURANTE, D. & DUNSON, D. B. (2022). Extended stochastic block models with application to criminal networks. *Annals of Applied Statistics* **16**.

- MÜLLER, P. AND QUINTANA, F. & ROSNER, G. L. (2011). A product partition model with regression on covariates. *J. Comput. Graph. Statist.* **20**.
- NOWICKI, K. & SNIJDERS, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *J. Amer. Statist. Assoc.* **96**.
- PITMAN, J. (2006). *Combinatorial Stochastic Processes*. Springer.
- RIGON, T., HSU, C.-L. & DUNSON, D. B. (2024+). A bayesian theory for estimation of biodiversity .
- SCHMIDT, M. N. & MORUP, M. (2013). Nonparametric bayesian modeling of complex networks: An introduction. *IEEE Signal Process. Mag.* **30**.
- SHARMA, H. & SRIVASTAVA, V. (2015). Descriptive analysis of social network measures. *International Journal of Modern Communication Technologies & Research (IJMCTR)* **3**.
- WADE, S. & GHAHRAMANI, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis* **13**.
- ZITO, A., RIGON, T. & DUNSON, D. B. (2024+). Bayesian nonparametric modeling of latent partitions via stirling-gamma priors .

Ringraziamenti

Ed ecco alla fine i tanto rinomati ringraziamenti, quella sezione che la maggior parte non vede l'ora di scrivere e invece io preferivo un altro capitolo di tesi. Infatti, mi ritrovo qua a scriverli un'ora prima di andare a stampare tutto (spoiler ci ho messo più di un'ora alla fine e rischiato che la copisteria chiudesse), perché dopo adesso non posso più rimandare...

Vorrei iniziare ringraziando il mio professore, Tommaso Rigon, per avermi proposto questo argomento che ha permesso di unire la mia passione per la criminologia con la statistica e per avermi seguito durante tutto questo percorso.

Grazie ai miei nonni, di Firenze e di Milano, che mi avete permesso di inseguire i miei obiettivi, consentendomi di trasferirmi a Milano. So che sono una nipote che si fa sentire poco, ma ciò non vuol dire che non vi pensi, sono solo fatta così.

Vorrei poi ringraziare la mia prima “casa”, Mamma, Lapo e Matteo. Come penso chiunque alla mia età mi sono trasferita felice e contenta verso la mia indipendenza, senza avere ripensamenti, ma la distanza mi ha fatto apprezzare di più i miei ritorni a casa, o i nostri ritrovarsi in qualche città diversa. Grazie soprattutto per non aver dato nuova vita alla mia cameretta e non averne preso possesso. Ah questo è dedicato solo per te Matte, smettila di crescere, che ogni volta che torno mi sento più bassa!!

...e la mia seconda “casa”, quella attuale, Ludo. Con te ho condiviso questi due anni di studio, nei nostri quaranta metri quadri. Non scriverò grazie per avermi supportato nelle mie notti insonni di studio, come suggeriva chat gpt, perché sai benissimo che non è vero, non perché te non mi avresti supportato, ma perché ho sempre dormito sogni tranquilli, per questo devi ringraziare te me, per aver trovato una studentessa poco ansiosa. Però mi sei sempre stato accanto nella quotidianità, portandomi sempre serenità e allegria, che forse è ancora più difficile.

Infine vorrei ringraziare tutte le mie amiche che mi sono state vicino in questi ultimi due anni, andrò in ordine temporale, così nessuno si offende...

E a proposito di tempo, bisogna ormai risalire a dieci anni fa per ricordare quando ci siamo conosciute, che vecchie che stiamo diventando... Avrete già capito che mi

sto riferendo a voi due, le mie compagne Superchicche (rimane per ora il costume per halloween più carino a cui abbiamo pensato, anche se il più bello rimangono sempre i punti neri ele occhiaie della Moni). Gli anni passano, le distanze tra di noi aumentano, le nostre vite cambiano, ognuna in una città diversa, ognuna con i suoi impegni (che brutto il mondo dei grandi, tornei volentieri ai nostri anni del liceo), le possibilità di vederci sono sempre meno, ma nonostante questo so che voi ci siete e ci sarete sempre. E visto che dite che non ve lo dico mai, ve lo scrivo direttamente, vi voglio bene!

Andando poi avanti di qualche anno, sono arrivati gli anni della triennale, dove ho conosciuto voi: Marti, Kri e Roberta. Ci siamo conosciute in anni un po' particolari, quelli del covid, ma nonostante tutte le restrizioni, quanto ci siamo divertite. Voi tutte trasferite a Firenze e io unica fiorentina che invece me ne sono andata, ognuna alla ricerca della propria strada. Siete molto diverse tra di voi, ma ognuna a suo modo mi ha insegnato tanto in questi anni. Marti con il tuo sorriso contagioso e con un pensiero sempre rivolto verso gli altri. Kri con la tua riservatezza, la tua calma e compostezza, che a volte puoi risultare un po' fredda, come penso un po' anche io, e forse per questo ci siamo sempre capite. E Roberta, con la tua voglia di vivere e quella giusta dose di sana pazzia, ma soprattutto con una lingua che ha addirittura meno peli della mia. Quando volete, ricordatevi che c'è sempre un divanetto pronto qua a Milano per voi.

E infine, ripeto non per importanza, è arrivata la magistrale qua a Milano e con lei sei arrivata te, Cami. Sei la prima persona con cui ho parlato il giorno del colloquio e rimanendo in tema statistica, quanto era bassa la probabilità che la prima persona con cui parlassi, fosse fuorisede come me, abitasse nel mio stesso isolato in una città grande come Milano e fosse pure simpatica! E invece è successo tutto questo, ci siamo conosciute e ci siamo prese fin da subito, rimanendo una al fianco dell'altra durante tutto questo percorso in una città nuova per entrambe. E nonostante ci vedessimo tutti i giorni e facessimo quasi tutto insieme, avevamo sempre qualcosa da dirci e raccontarci, tanto da fare nottata solo io e te, Io aspetto sempre il tuo ritorno qua a Milano e il gossip delle aziende milanesi aspetta solo te e le nostre future pause pranzo.