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

Scienze Statistiche

Tesi di Laurea

Analisi della struttura di vendita di una compagnia assicurativa: un modello bayesiano non parametrico per una rete di reti

Relatore Prof. Bruno Scarpa
Dipartimento di Scienze Statistiche

Correlatore Dott. Daniele Durante

Laureanda Sally Paganin
Matricola N 1043401

Anno Accademico 2014/2015

Indice

1 La struttura di vendita in termini di rete **1**
1.1 I dati . . . 1
1.2 Definizione di rete . . . 3
1.3 La rete di vendita . . . 4

2 Analisi in termini di rete sociale **9**
2.1 Misure di rete locali . . . 11
2.2 Misure di rete globali . . . 12
2.3 I modelli classici di rete . . . 14
2.4 Modelli a blocchi stocastici . . . 18

3 Un modello bayesiano per una rete di reti **21**
3.1 Modello di clustering per le reti di secondo livello . . . 22
3.1.1 Distribuzione a priori . . . 24
3.1.2 Distribuzione a posteriori . . . 26
3.1.3 Gibbs Sampling . . . 28
3.2 Modello a blocchi stocastici per la rete di primo livello . . . 30
3.2.1 Stima del modello . . . 31

4 Studio di simulazione **33**
4.1 Simulazione di una rete di reti . . . 33
4.2 Stima e *label-switching* . . . 35

5 Applicazione ai dati **37**
5.1 Discussione sulla stima del modello . . . 37
5.2 Risultati . . . 38

Introduzione

La dinamica di vendita delle compagnie assicurative si differenzia in maniera particolare da quella caratterizzante aziende di altri settori, in quanto il rapporto con la clientela non è gestito in maniera diretta, ma è intermediato da un complesso di attività disseminate sul territorio, le *agenzie assicurative*. Esse costituiscono il principale canale di vendita, e sono delle strutture imprenditoriali, formate da uno o più soci, finalizzate alla gestione e acquisizione degli affari assicurativi. L'organizzazione e l'operatività di una agenzia sono caratterizzate da una serie di linee guida generali concordate con la compagnia con la quale ha sottoscritto il mandato, e da un insieme di figure professionali che collaborano con gli agenti quali subagenti, produttori, impiegati di agenzia. L'*agente* è la figura che generalmente interagisce con il cliente, con l'obiettivo di far sottoscrivere il prodotto assicurativo adatto alle esigenze del cliente stesso. Per esercitare questa attività di intermediazione gli agenti devono essere iscritti ad un apposito registro (RUI, Registro Unico Intermediari).

Il ruolo della compagnia mandante è quello di mettere a disposizione i prodotti assicurativi, e dare obiettivi ed incentivi sui volumi di vendita; inoltre agisce sul proprio mercato con operazioni pubblicitarie e soprattutto attraverso azioni di marketing rivolte alle agenzie (*trade marketing*). La relazione tra compagnia e agenzia non è tuttavia univoca, nel senso che un'agenzia può proporre e gestire contratti relativi a prodotti assicurativi anche di compagnie diverse; le agenzie sono quindi entità dotate di una propria autonomia decisionale per quanto riguarda la politica di vendita.

Un passo fondamentale per lo sviluppo di campagne di *trade marketing* efficaci e di impatto consiste nel mettere in evidenza le caratteristiche delle agenzie e delle relative politiche di vendita. Una campagna ideale dovrebbe essere in grado di tenere conto delle politiche di vendita di ogni singola agenzia e in base a queste, differenziare le strategie di marketing in modo da massimizzare il profitto, ma comporta un costo troppo elevato per un'azienda in termini economici e di risorse umane. D'altra parta una campagna unica per tutte le agenzie è certamente meno dispendiosa, ma rischia di non portare comunque ad un profitto conveniente. Una soluzione di compromesso consiste nel definire degli insiemi di agenzie per le quali è possibile attivare una stessa campagna; a tale scopo, nel presente elaborato si propone un'analisi alternativa del paradigma di vendita assicurativo secondo strutture di rete, ovvero strutture costituite da un insieme di attori e di relazioni tra gli stessi, che possono essere ricondotte ad un formalismo matematico usando la teoria dei grafi.

L'*analisi delle reti sociali* (SNA, dal termine inglese *Social Network Analysis*) definisce l'insieme di metodologie e modelli volto allo studio strutture relazionali. La SNA rappresenta

un'area di ricerca sempre più popolare, in quanto trova applicazioni in svariate aree di studio anche molto differenti tra loro, quali ad esempio le scienze sociali, la biologia o le neuroscienze. La definizione di modelli statistici *ad hoc* per dati di rete è motivata dal fatto che i metodi statistici classici non risultano efficaci nel descrivere dati di rete, in quanto non sono in grado di cogliere le strutture di dipendenza tra gli attori.

Nel Capitolo 1 si descrive in maniera più approfondita la rappresentazione del paradigma di vendita in termini di reti e si presentano i dati disponibili; il Capitolo 2 si discutono alcune tra le più comuni metodologia di analisi di rete, mentre nel Capitolo 3 si presenta un approccio alternativo basato su un modello bayesiano non parametrico per una rete di reti. Il modello è valutato nel Capitolo 4 attraverso uno studio simulazione, mentre il Capitolo 5 riporta i risultati ottenuti.

Capitolo 1

La struttura di vendita in termini di rete

1.1 I dati

I dati analizzati nella seguente trattazione sono stati forniti da una compagnia assicurativa italiana, e sono stati elaborati con l'aiuto di un esperto del settore. Delle agenzie affiliate alla compagnia, è stato selezionato solamente un campione di agenzie di interesse, in base alla loro inclusione a 4 particolari progetti. Sono disponibili i dati di portafoglio relativi a 135 agenzie, reperiti a termine dell'anno 2014; per ogni agenzia si hanno tutte le polizze stipulate dalla stessa e, per ogni polizza, il prodotto alla quale è associata e il cliente che l'ha stipulata. Le compagnie assicurative in genere possiedono un numero di prodotti decisamente elevato, e per questa compagnia specifica erano disponibili più di 200 prodotti; ai fini dell'analisi, si è deciso di raggrupparli in 15 tipologie di prodotti, di seguito elencate:

- Casa
- Attività commerciali
- Auto
- Auto R.D. (Rischi Diversi)
- Altri danni
- RC (Rischi Civili)
- Credito e cauzioni
- Infortuni
- Investimento
- Malattia
- Previdenza

- Protezione
- Risparmio
- Vita collettive
- Altro

Si può pensare al portafoglio clienti di una compagnia assicurativa, come la somma dei portafogli delle singole agenzie, i quali variano sia a seconda delle caratteristiche delle stesse (dimensione dell'agenzia, territorio in cui si trova, etc), che delle politiche di vendita da esse adottate. Per dare un'idea delle variabilità e della grandezza dei portafogli, in Figura 1.1 sono riportate rispettivamente i conteggi di frequenza del numero di clienti e del numero di polizze per agenzia. Si può notare in maniera immediata la presenza nel campione di una ventina di agenzie di piccole dimensioni, in quanto hanno un numero di clienti/polizze sull'ordine della decina di unità. Ovviamente vi è una correlazione tra le due numerosità, dato che ad un numero maggiore di clienti corrisponde un maggiore numero di polizze; tra l'altro le distribuzioni presentano un andamento simile, con la differenza che il numero di polizze ha un campo di variazione che è circa il doppio di quello dei clienti; come si può notare dal diagramma di dispersione Figura 1.1 il rapporto tra polizze e clienti è maggiore di 1. Infatti questi ultimi hanno la possibilità di stipulare più polizze relative ad uno o più prodotti, ed è possibile operare una distinzione tra clienti *mono-prodotto* e clienti *pluri-prodotto* sulla quale definire la politica di vendita dell'agenzia.

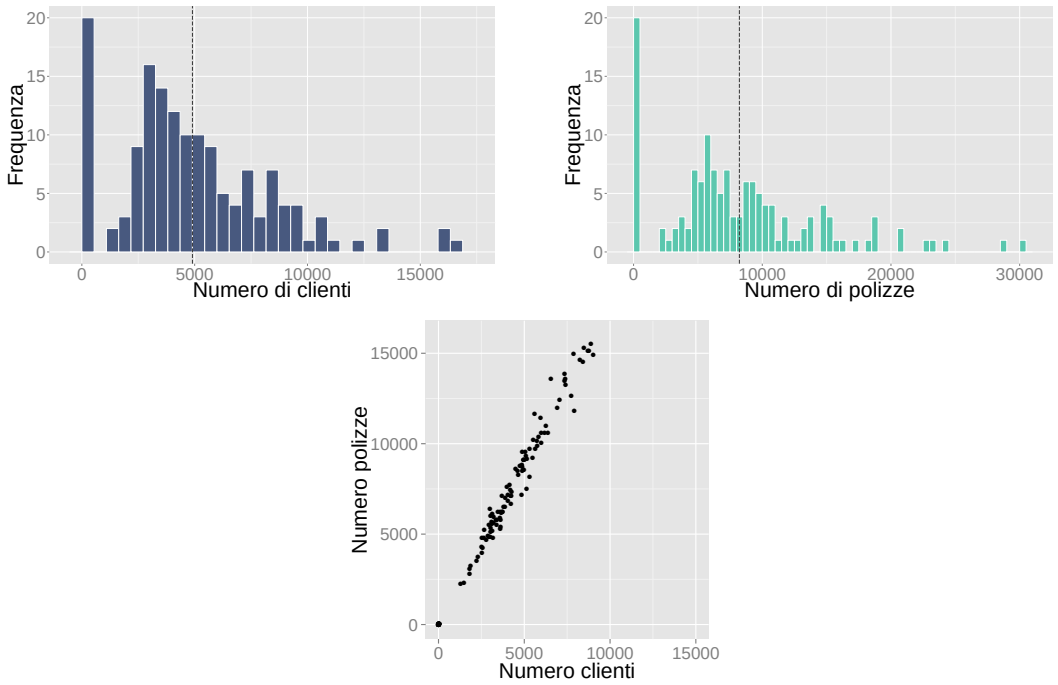

Figura 1.1: Prima riga: frequenze del numero di clienti e del numero di polizze relative ai dati di portafoglio campione di agenzie. La linea tratteggiata indica il valore medio, 4580 clienti (s.d. 3300) e 8220 polizze (s.d. 6106). Seconda riga: diagramma di dispersione del numero di clienti e di polizze per le reti; ad un cliente corrispondono all'incirca due polizze

Una delle strategie principali messe in atto dalle aziende di servizi per l'acquisizione di nuovi contratti, consiste nel proporre ulteriori prodotti ai clienti già acquisiti. Essi rappresentano la più probabile e sicura fonte di un nuovo contratto, in quanto hanno già una relazione con l'agente che va a vantaggio dello stesso: è più semplice per un agente assicurativo interagire con clienti che già lo conoscono e dei quali ha conquistato la fiducia nell'arco del tempo, piuttosto che trovarne di nuovi e costruire da zero relazioni con essi. Inoltre la conoscenza del cliente facilita l'individuazione di nuovi bisogni da poter soddisfare con la proposta di un nuovo contratto, mentre dall'altro lato, il cliente risulta più propenso ad accettarlo, data la relazione di fiducia già creatasi in precedenza.

L'ufficio marketing della compagnia assicurativa vuole proporre una campagna di per incentivare la nascita di clienti multi-prodotto; a tale scopo è di interesse non solo conoscere quali sono i tipi di prodotti che vengono sottoscritti più spesso da uno stesso cliente, ma definire l'associazione tra prodotti in termini probabilistici. Si può presumere che all'interno del portafoglio di ogni agenzia, vi sia una distribuzione differente delle tipologie di cliente pluri-prodotto, e che sia influenzata della politica di vendita adottata dalla stessa; pensare però di proporre un'azione di marketing differente per ogni agenzia richiede un impiego di risorse troppo dispendioso per l'azienda. Per limitare l'impiego di risorse una possibilità è quella di definire degli insiemi di agenzie simili tra loro in termini di portafoglio e di politica di vendita, e proporre quindi una campagna specifica per ogni insieme. Dato questo obiettivo, si vogliono definire degli insiemi di agenzie simili tra loro in termini di portafoglio, e indagare se risultano simili anche in termini di politica di vendita; in altre parole, se agenzie con le stesse associazioni tra le tipologie di prodotto risultano avere anche la stessa distribuzione di prodotti.

Si è pensato di analizzare il paradigma di vendita attraverso l'uso dei grafi, in quanto le associazioni tra le tipologie di prodotti sono per natura di tipo binario e dipendenti tra loro: è presumibile che un cliente che abbia sottoscritto più polizze con la stessa compagnia sia propenso a sottoscriverne una terza, ma la tipologia di prodotto dipenderà dalle due che già possiede. Si può definire quindi un'insieme di reti di prodotti che descrivono le politiche di vendita delle agenzie, le quali possono a loro volta essere inserite all'interno di un contesto di rete atto a definire delle relazioni di similarità tra di esse. Si riporta di seguente la definizione di rete in termini matematici.

1.2 Definizione di rete

Una rete è un "insieme di oggetti interconnessi"[1] usualmente rappresentata per mezzo di un grafo. Formalmente, un grafo $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ è definito da un insieme $\mathcal{V}$ di *vertici* (o *nodi*) e un insieme $\mathcal{E}$ di *archi*, i cui elementi sono coppie di vertici distinti $\{u, v\}, u, v \in \mathcal{V}$. Le cardinalità di tali insiemi, ovvero il numero di vertici $V = |\mathcal{V}|$ e il numero di archi $E = |\mathcal{E}|$, definiscono rispettivamente l'*ordine* e la *grandezza* del grafo $\mathcal{G}$. Ad un grafo di questo tipo è usualmente

[1] Oxford Dictionary

associata una rappresentazione matriciale $\boldsymbol{A}$ di dimensione $V \times V$ che descrive la presenza o meno di una relazione tra i nodi, e i cui elementi sono definiti:

$$a_{ij} = \begin{cases} 1 & \text{se } \{i,j\} \in E \\ 0 & \text{altrimenti} \end{cases} \tag{1.1}$$

Un grafo di tipo *orientato* (detto anche *digrafo*) è definito in modo analogo, ma gli archi sono delle coppie ordinate di vertici e sono anch'essi detti *orientati*; un grafo è invece *pesato* se ad ogni arco è associato un valore numerico. Un arco $\{u, v\}$ con $u = v$, è detto *self-loop* (cappio), mentre un *arco multiplo* è un arco che compare più di una volta; nella presente tesi si considerano solamente *grafi semplici*, ovvero grafi che non contengono né cappi né archi multipli, di tipo *non orientato* e non pesato.

1.3 La rete di vendita

Allo scopo di ridefinire il paradigma di vendita della compagnia assicurativa in termini di reti, è stato necessario determinare quali fossero gli attori principali tra le entità in esame, e quali fossero le relazioni tra di essi. Nella prima parte della Figura 1.2 si ha una rappresentazione della struttura di vendita: la compagnia gestisce un insieme di agenzie, ognuna delle quali tiene rapporti con un certo numero di clienti che hanno sottoscritto contratti per uno o più prodotti. Si suppone per semplicità che la relazione tra cliente e agenzia sia univoca, ovvero che un cliente faccia riferimento ad una sola e unica agenzia.

Nel secondo riquadro della Figura 1.2 si mostra come le entità coinvolte nel processo di vendita possano essere collocate all'interno di una struttura di rete che si articola in due livelli. Il primo livello rappresenta le agenzie come attori di una rete che hanno tra loro una relazione basata sulla similarità mentre il secondo livello ha lo scopo di rappresentare la politica di vendita di ogni agenzia. Si ha quindi associata ad ogni agenzia, una rete in cui i nodi sono i prodotti e la relazione tra essi è basata sui clienti.

Formalizzando in termini matematici, sia $\mathcal{G}$ un grafo non orientato su un insieme di nodi $\mathcal{N}$ con cardinalità $|\mathcal{N}| = N$, definita come *rete di primo livello* (rete delle agenzie) con matrice di adiacenza X di elementi $x_{ij} \in \{0, 1\}$ che codificano la presenza o meno di un arco tra i nodi. Ad ogni nodo i, $i = 1, \ldots, N$, è a sua volta associato un'ulteriore grafo $\mathcal{H}_i$; ogni grafo è definito su uno stesso insieme di nodi $\mathcal{V}$ (i prodotti) ma è caratterizzato da connessioni diverse, descritte da una matrice di adiacenza denotata con A_i. Definiamo l'insieme delle reti associate ai nodi, come *reti di secondo livello.*

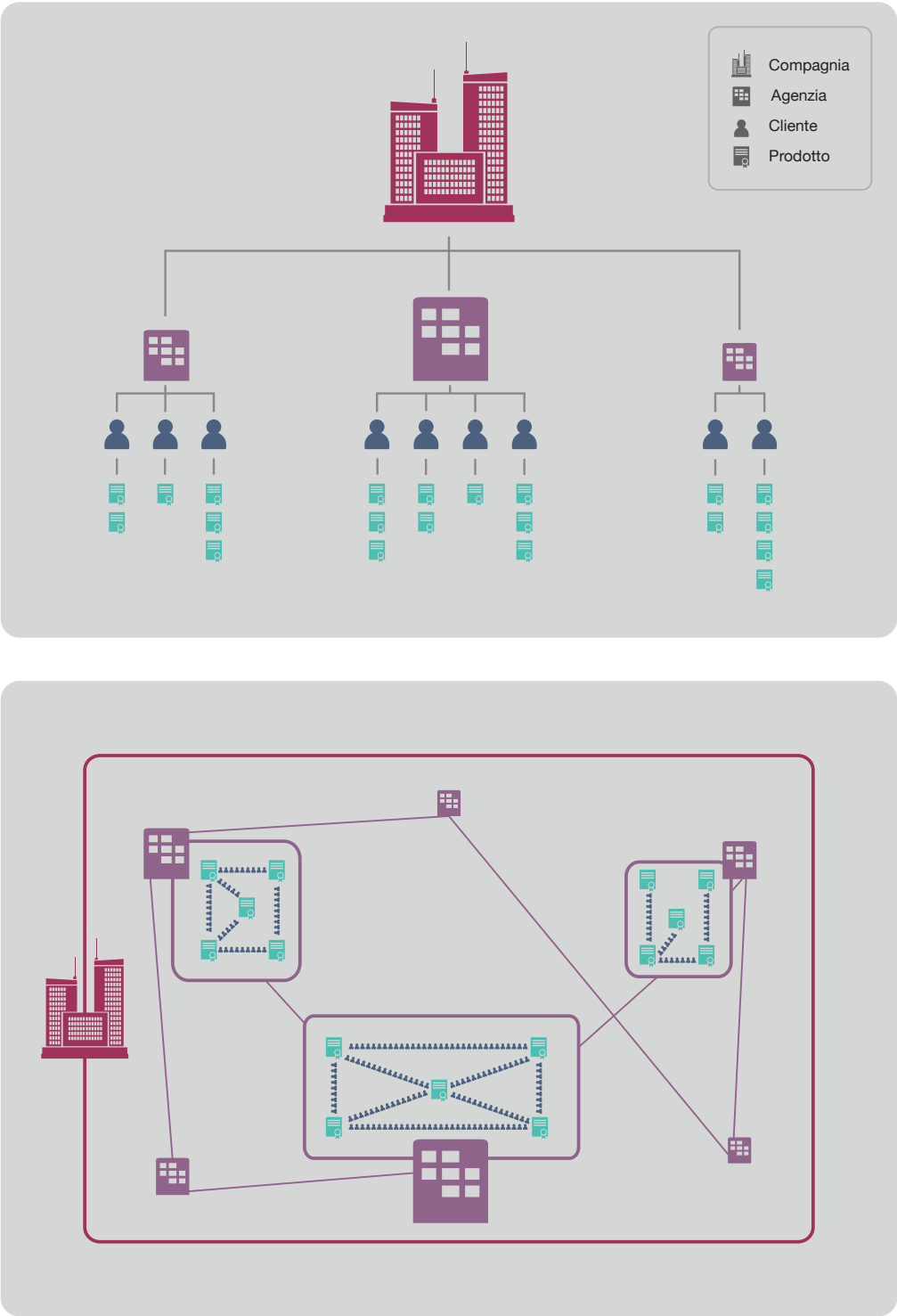

Figura 1.2: Rappresentazione della struttura di vendita di una compagnia assicurativa e ridefinizione della struttura in termini di rete. La compagnia controlla le agenzie, le quali gestiscono un insieme di clienti che stipulano polizze per uno o più prodotti assicurativi. In termini di rete, le agenzia sono tra loro collegate secondo una relazione di amicizia basata sulla similarità; ad ogni agenzia è associata una rete che descrive il portafoglio della clientela associata. In particolare i prodotti rappresentano i nodi della rete mentre i clienti pluri-prodotto definiscono le relazioni tra di essi; in questo modo i nodi sono comuni, mentre le connessioni tra di essi variano a seconda dell'agenzia.

Decisi gli attori delle reti, è necessario definire le relazioni tra di essi al fine di rappresentare al meglio la struttura di vendita. Per quanto riguarda le agenzie si è scelto di selezionare delle variabili che le descrivessero e di definire una relazione basata sulla similarità: agenzie ch si assomigliano abbastanza tra loro sono quindi considerate connesse. In particolare si è considerata la distribuzione percentuale delle 15 tipologie di prodotti assieme ad altre 4 variabili dicotomiche descriventi le attività dell'agenzia stessa quali:

- P1 = 1, se l'agenzia è stata inclusa in un progetto di CRM (*Customer Relationship Mangement*), per la fidelizzazione del cliente.
- P2 = 1, se l'agenzia è stata selezionata per un progetto di marketing attivo.
- P3 = 1, se l'agenzia fa parte di un organo organizzativo detto *direttivo*.
- P4 = 1, se l'agenzia fa parte di una particolare regione italiana.

Un'agenzia può essere caratterizzata da più di un attività: si riporta in Figura 1.3 un diagramma di Venn riportante il numero di agenzie per intersezione delle variabili.

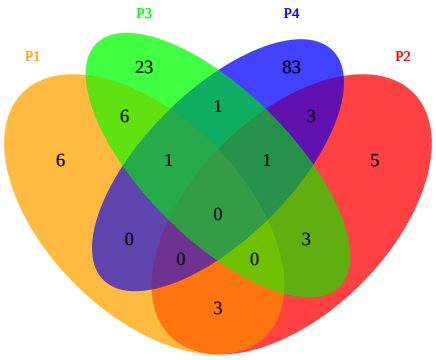

Figura 1.3

Avendo a disposizione sia variabili di tipo quantitativo che qualitativo, si è utilizzato come indice di dissimilarità, la distanza di Gower (Gower, 1971), definita come la somma pesata delle distanze tra gli individui, calcolate variabile per variabile, tenendo conto della tipologia:

$$d(i,j) = \frac{\mathrm{w}_1\delta(i,j)d(i,j;1) + \ldots + \mathrm{w}_p\delta(i,j;p)d(i,j;p)}{\sum_{k=1}^{p} \mathrm{w}_k\delta(i,j;k)} \tag{1.2}$$

In altre parole, $d(i,j)$ non è altro che la media pesata delle distanze $d(i,j,k)$, con pesi $w_k\delta(i,j;k)$. La quantitè $\delta(i,j;k)$ è pari a 0 nel caso in cui la k-esima variabile risulti mancante per almeno uno dei due individui, oppure nel caso la variabile sia di tipo binario asimmetrico ed entrambi i valori per gli individui siano pari a 0. In tutti gli altri casi $\delta(i,j;k)$ è pari a 1. Il contributo $d(i,j;k)$ alla distanza totale di una variabile di tipo qualitativo è pari a 0 se gli individui hanno entrambi lo stesso valore, altrimenti è pari a 1; il contributo delle altre

variabili è calcolato invece come la differenza assoluta dei valori dei due individui, divisa per l'intervallo di variazione totale della variabile.

Volendo ottenere una rappresentazione binaria della relazione tra le agenzie, si è scelta come soglia di connessione la media complessiva delle distanza: due agenzie sono dunque connesse se la loro distanza è minore della distanza media. Si noti che in una rappresentazione di rete, non è necessario che le connessioni siano di tipo binario, ma è possibile considerare delle relazione pesate, ad esempio come $1 - d(i, j)$: agenzie vicine in termini di similarità avranno una relazione più forte rispetto a quella tra agenzie più distanti tra loro. In questa trattazione ci riserviamo di considerare solo relazioni tra nodi di tipo binario.

Ad ogni agenzia è associata una rete di prodotti, definita allo scopo di caratterizzare la tipologia di clienti multi-prodotto, e quindi la politica di vendita, di ognuna di esse. La relazione tra i prodotti è stata quindi basata sui clienti nel seguente modo: per ogni coppia di prodotti si è calcolata la proporzione di clienti che possiedono entrambi i prodotti rispetto il totale dei clienti che possiedono almeno uno dei due. Definito $\#P_i$ come l'insieme di clienti che hanno sottoscritto un polizza per il prodotto i, si è calcolata una matrice 15×15 contenente le seguenti quantità:

$$\frac{\#P_i \bigcap \#P_j}{\#P_i \bigcup \#P_j}, \quad i = 1, \ldots, 15, \quad j = 1, \ldots, 15 \tag{1.3}$$

Volendo definire una relazione binaria per le connessioni, si è definita una soglia con il seguente metodo. Fissato una valore sulla distribuzione delle quantità calcolate, si sono ottenute le relative reti di prodotti, e si è utilizzato un metodo di *clustering* gerarchico di tipo completo sulle reti ottenute, basato sempre sulla distanza di Gower, che per variabili di tipo qualitativo è definita come sopra. Si sono ottenuti 6 gruppi di reti di prodotti e calcolato un indice di correlazione di rete, tra gli indici dei gruppi ottenuti e la rete di agenzie, definita come *assortatività* (Newman, 2003):

$$r = \frac{\sum_i e_{ii} - sum_i a_i b_i}{1 - sum_i a_i b_i} \tag{1.4}$$

dove e_{ii} è la proporzione di archi che connettono i vertici di tipo i e j, mentre $a_i = \sum_j e_{ij}$ e $b_j = \sum_i e_{ij}$. Si è quindi scelto il valore che massimizza tale quantità, ottenendo come soglia di connessione il quantile 0.25 della distribuzione di tali quantità per ognuna delle reti.

In Figura 1.4 è rappresentata la rete delle agenzie ottenuta, identificata con un numero da 1 a 135, in cui i nodi sono posizionati secondo l'algoritmo di Fruchterman & Reingold (1991); per alcune agenzie selezionate, in Figura 1.5 si ha una rappresentazione delle matrici di adiacenza relative alle reti di prodotti: ad ogni quadrato corrisponde un coppia di prodotti, che risulta colorato se sono connessi.

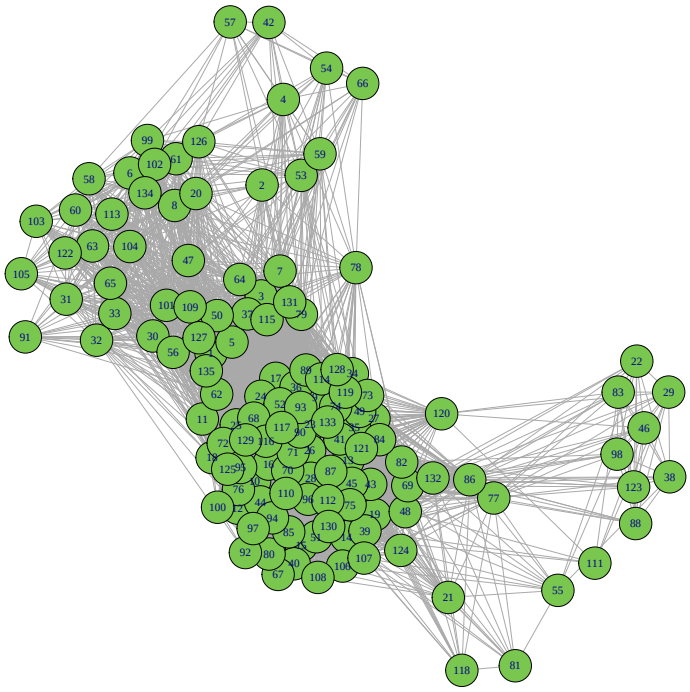

Figura 1.4: Rappresentazione della rete di agenzie

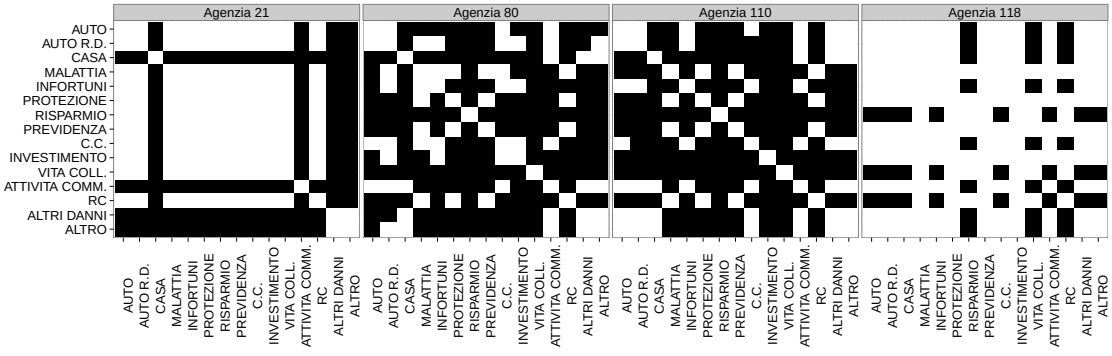

Figura 1.5: Per alcune agenzie selezionate: rappresentazione delle matrici di adiacenza delle relative reti di prodotti. Ad ogni quadrato corrisponde una coppia di prodotti; se il quadrato è colorato i due prodotti sono connessi, altrimenti è lasciato in bianco

Capitolo 2

Analisi in termini di rete sociale

La definizione di modelli statistici *ad hoc* per dati di rete è motivata dal fatto che i metodi statistici classici non risultano efficaci nel descrivere dati di rete, in quanto non sono in grado di cogliere le strutture di dipendenza tra gli attori. Si consideri un approccio di tipo classico al problema. Una strategia di analisi è quella di trattare i dati come in un contesto di classificazione di tipo binario: le connessioni tra i nodi della rete di primo livello, ovvero le agenzie, sono la variabile dipendente, mentre le connessioni tra i prodotti costituiscono l'insieme di variabili esplicative; ad ogni coppia di agenzie, è associata una coppia di vettori i cui elementi sono tutte le possibili connessioni tra i prodotti (se $V = 15$ è il numero di prodotti, si hanno $V(V-1)/2 = 105$ possibili coppie). È plausibile che due agenzie simili in termini di connessioni di prodotti, abbiano una maggiore probabilità di essere tra loro collegate. Si definisce un nuovo vettore di variabili esplicative, come misura di similarità tra le due reti di prodotto secondo la concordanza delle variabili: se l'arco tra due nodi prodotto è assente o presente in entrambe le reti, la nuova variabile è pari a 1, altrimenti 0. Si ottiene quindi una riduzione del modello ad un problema di classificazione binaria, in cui anche le variabili esplicative hanno un carattere dicotomico, che si può affrontare facendo uso degli usuali modelli di *data mining* (si veda Azzalini & Scarpa (2012) e Hastie et al. (2009) per una rassegna dei metodi). Tuttavia un approccio simile suppone che le variabili esplicative siano indipendenti tra loro, mentre in presenza di dati di tipo relazionale, vi sono delle dipendenze strutturali tra le variabili. A titolo di esempio si riportano in Tabella 2.1 i risultati della stima di un modello lineare generalizzato ottenuto tramite una procedura di tipo *forward stepwise*, che seleziona 46 variabili come significative, delle 105 disponibili:

Un modello simile ha il vantaggio di offrire una facile interpretazione in termini di *odds-ratio*, ma si adatta in maniera povera ai dati: considerando ogni variabile come indipendente, non è possibile cogliere le strutture di dipendenza all'interno delle reti, in quanto la presenza di un arco tra due nodi non dipende solamente dagli attributi ad essi associati, ma dal complesso delle interazioni che caratterizzano le reti.

Per riuscire a utilizzare appieno l'informazione relativa alla struttura delle reti di secondo livello, è necessario ottenere una rappresentazione ridotta delle loro strutture di dipendenza. Un approccio di base è quello di utilizzare le misure di rete derivanti dalla teoria dei grafi, come descrittive delle reti di prodotti, e includere tali misure come covariate nei modelli

	Modello GLM
(Intercetta)	−3.95 (0.31)***
‘CASA – INFORTUNI‘	0.26 (0.13)*
‘INFORTUNI – PROTEZIONE‘	0.10 (0.13)
‘CASA – AUTO R.D.‘	0.36 (0.10)***
‘INVESTIMENTO – RISPARMIO‘	1.08 (0.14)***
‘AUTO R.D. – INFORTUNI‘	−3.28 (0.26)***
‘AUTO R.D. – MALATTIA‘	0.34 (0.09)***
‘AUTO – AUTO R.D.‘	1.52 (0.19)***
‘PROTEZIONE – VITA COLLETTIVE‘	0.33 (0.14)*
‘CASA – MALATTIA‘	1.56 (0.19)***
‘ATTIVITA COMMERCIALI – MALATTIA‘	−1.06 (0.12)***
‘ALTRI DANNI – MALATTIA‘	1.29 (0.18)***
‘ALTRI DANNI – VITA COLLETTIVE‘	−0.79 (0.31)*
‘AUTO – RISPARMIO‘	2.17 (0.35)***
‘CREDITO E CAUZIONI – INVESTIMENTO‘	0.34 (0.08)***
‘AUTO R.D. – CREDITO E CAUZIONI‘	0.22 (0.05)***
‘ALTRI DANNI – PREVIDENZA‘	−0.31 (0.26)
‘PROTEZIONE – RISPARMIO‘	0.78 (0.15)***
‘ALTRI DANNI – ATTIVITA COMMERCIALI‘	0.09 (0.07)
‘AUTO – RC‘	−0.85 (0.23)***
‘ALTRO – PROTEZIONE‘	1.05 (0.30)***
‘CREDITO E CAUZIONI – RC‘	0.27 (0.13)*
‘MALATTIA – PREVIDENZA‘	1.74 (0.40)***
‘ALTRI DANNI – PROTEZIONE‘	−1.49 (0.43)***
‘RC – RISPARMIO‘	−0.83 (0.18)***
‘ALTRI DANNI – INVESTIMENTO‘	0.51 (0.13)***
‘AUTO – PROTEZIONE‘	0.24 (0.24)
‘AUTO – PREVIDENZA‘	−0.75 (0.21)***
‘CASA – CREDITO E CAUZIONI‘	0.27 (0.10)**
‘CASA – ALTRO‘	−0.09 (0.26)
‘CASA – ATTIVITA COMMERCIALI‘	0.61 (0.13)***
‘ALTRI DANNI – RISPARMIO‘	−1.37 (0.44)**
‘CREDITO E CAUZIONI – VITA COLLETTIVE‘	−0.70 (0.29)*
‘ALTRO – CREDITO E CAUZIONI‘	0.96 (0.33)**
‘ALTRI DANNI – AUTO R.D.‘	−0.61 (0.21)**
‘AUTO R.D. – PROTEZIONE‘	0.95 (0.22)***
‘CREDITO E CAUZIONI – PROTEZIONE‘	−0.83 (0.22)***
‘ALTRI DANNI – RC‘	1.28 (0.44)**
‘CREDITO E CAUZIONI – PREVIDENZA‘	0.12 (0.05)*
‘CREDITO E CAUZIONI – MALATTIA‘	0.23 (0.09)*
‘ALTRO – INVESTIMENTO‘	−0.38 (0.11)***
‘ATTIVITA COMMERCIALI – INVESTIMENTO‘	0.33 (0.10)***
‘RISPARMIO – VITA COLLETTIVE‘	−0.15 (0.07)*
‘CASA – ALTRI DANNI‘	−0.68 (0.30)*
‘ALTRO – AUTO‘	0.81 (0.33)*
‘ALTRI DANNI – CREDITO E CAUZIONI‘	−0.43 (0.23)
AIC	9646.38
BIC	9973.44
Log Likelihood	−4777.19
Deviance	9554.38
Num. obs.	9045

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Tabella 2.1: Coefficienti stimati per il modello GLM, con relativo errore standard tra parentesi.

per dati di rete. Di seguito si presentano alcune misure utilizzate comunemente nell'analisi descrittiva di rete, facendo distinzione tra misure di tipo *locale* e misure *globali*, e se ne riportano i risultati per le reti di secondo livello.

2.1 Misure di rete locali

Le misure di rete di tipo *locale* rappresentano quell'insieme di statistiche atte a caratterizzare i singoli nodi all'interno della rete. Si riportano nel seguito le definizioni di quelle usate nelle analisi.

- Grado: dato un nodo v, il grado corrisponde al numero di archi incidenti ad esso, e si può esprimere formalmente utilizzando la matrice di adiacenza: $d_v = \sum_{j=1}^{V} a_{vj}$.
- Eccentricità: si tratta di una misura di nodo, $e(v)$ basata sul concetto di *distanza* tra nodi: la distanza $d(u, v)$ tra due nodi $\{u, v\} \in V$ è detta *distanza geodetica*, e corrisponde alla lunghezza del cammino minimo tra i due nodi[1]. L'eccentricità corrisponde alla più grande distanza tra un nodo v ed ogni altro nodo $u \in V \backslash \{v\}$: $e(v) = max_{v \in \mathcal{V}} d(u, v)$
- Misure di centralità: definiscono l'importanza di un nodo all'interno della rete; sono state proposte diverse misure nel corso degli anni, che si differenziano in base a quale criterio viene considerato nel definire un nodo come "importante". Nel seguito consideriamo le più utilizzate:
  - centralità *betweenness*: misura quanto un nodo si trova "in mezzo" ad altri due nodi; l'importanza di un nodo è determinata in questo caso, in base alla sua posizione all'interno di un cammino. La misura più comunemente utilizzata è quella introdotta da Freeman (1977) definita come

$$C_b(v) = \sum_{s \neq t \neq v \in \mathcal{V}} \frac{\sigma(s, t|v)}{\sigma(s, t)} \tag{2.1}$$

    dove $\sigma(s, t|v)$ è il numero di cammini minimi tra i vertici s e t che passano per v, mentre $\sigma(s, t)$ è il numero totale di cammini minimi tra s e t. È possibile normalizzare tale misura nell'intervallo $[0, 1]$, moltiplicandola per un fattore pari a $(N-1)(N-2)/2$, dove N è il numero di nodi nel grafo, in modo da permettere il confronto con le altre misure.
  - centralità *closeness* (Sabidussi, 1966): definisce in quale misura un nodo è "vicino" a tutti gli altri, ed è data dall'inverso della somma delle distanza geodetiche:

$$C_{cl}(v) = \frac{1}{\sum_{v \neq u} d(u, v)} \tag{2.2}$$

    La versione normalizzata nell'intervallo $[0, 1]$, prevede la moltiplicazione per un fattore pari a $N - 1$.

[1] Nella teoria dei grafi il cammino minimo, è il percorso che collega due vertici dati e che minimizza la somma dei costi data dall'attraversare ciascun arco

– centralità basata sugli autovalori (*eigenvector*): si basa sul concetto di "prestigio" di un nodo, nel senso che maggiormente centrali sono i vicini di un nodo, più lo è il nodo stesso. Questo genere di misura è tipicamente espressa come soluzione in termini di autovettori, di un sistema di equazioni lineari.

$$C_{Ei}(v) = \alpha \sum_{\{u,v\} \in \mathcal{E}} c_{Ei}(u) \tag{2.3}$$

Il vettore $\boldsymbol{C}_{Ei} = (C_{Ei}(1), \ldots, C_{Ei}(N))^T$ rappresenta la soluzione all'equazione $\boldsymbol{A}\boldsymbol{C}_{Ei} = \alpha^{-1}\boldsymbol{C}_{Ei}$, dove $\boldsymbol{A}$ è la matrice di adiacenza associata al grafo $\mathcal{G}$; una scelta valida per α^{-1} è data dal maggiore degli autovalori relativi alla matrice $\boldsymbol{A}$, e quindi $\boldsymbol{C}_{Ei}$ risulta essere il corrispondente autovalore (Bonacich, 1972).

- Transitività locale: è anche definita coefficiente di *clustering* e corrisponde al numero di triangoli (insieme di tre vertici connessi tra loro) contenenti il nodo v in rapporto al numero di possibili triangoli centrati su quel nodo. Preso un nodo di interesse, tale misura può essere interpretata come la probabilità di avere un arco tra due nodi vicini estratti casualmente.

$$C(v) = \frac{|e_{vw} \in \mathcal{E} : v, w \in N(v)|}{k_i(k_i - 1)/2} \tag{2.4}$$

dove $N(v)$ indica l'insieme di nodi vicini a v.

Nella Tabella 2.2 si riportano le statistiche delle distribuzioni delle misure sopra descritte per i nodi-prodotto in tutte le reti di secondo livello. Si può notare una discreta variabilità di tutte le misure tra le reti; ad esempio, tutti i nodi presentano un grado mediamente elevato, sintomo che vi è un buon numero di connessioni all'interno delle reti, ma la composizione delle connessioni risulta piuttosto variabile. Anche le misure di transitività ed eccentricità delle reti mostrano che i nodi sono generalmente ben collegati tra loro: la distanza massima dei nodi tra loro è pari a 2, mentre il coefficiente di *clustering* è mediamente ungule per tutti i nodi. Per quanto riguarda la centralità dei nodi, i tipi di prodotto Vita Collettive e Investimento, sembrano essere quelli più importanti, ma anche quelli con variabilità maggiore.

2.2 Misure di rete globali

Le misure globali di rete sono misure che descrivono l'intera struttura di dipendenza della rete nel suo complesso.

- Densità: corrisponde al rapporto tra il numero di archi presenti in un grafo e il numero di possibili connessioni; varia tra 0 (nessun arco nel grafo) e 1 (tutti i nodi sono tra loro connessi).

- Diametro e raggio: sono misure definite sulla distanza geodetica. Il diametro corrisponde all'eccentricità massima all'interno della rete, ovvero alla distanza massima tra due nodi; al contrario, il raggio è dato dall'eccentricità minima della rete.

	Grado	Betweenness	Closeness	Autovalori	Eccentricità	Transitività locale
Casa	11.00 (3.15)	0.04 (0.10)	0.80 (0.18)	0.85 (0.14)	1.84 (0.46)	0.64 (0.23)
Altri Danni	10.00 (3.38)	0.02 (0.02)	0.75 (0.18)	0.77 (0.18)	1.89 (0.42)	0.66 (0.22)
Altro	9.00 (2.99)	0.01 (0.02)	0.71 (0.16)	0.71 (0.16)	1.90 (0.40)	0.65 (0.22)
Attività Commerciali	9.00 (3.13)	0.01 (0.02)	0.72 (0.17)	0.73 (0.17)	1.89 (0.42)	0.66 (0.22)
Auto	9.00 (3.12)	0.01 (0.01)	0.72 (0.16)	0.74 (0.17)	1.90 (0.40)	0.69 (0.23)
Auto R.d.	8.00 (2.69)	0.01 (0.01)	0.68 (0.15)	0.67 (0.14)	1.93 (0.38)	0.67 (0.23)
Credito E Cauzioni	9.00 (2.76)	0.01 (0.01)	0.69 (0.15)	0.68 (0.15)	1.93 (0.38)	0.67 (0.23)
Infortuni	11.00 (3.40)	0.02 (0.01)	0.77 (0.17)	0.81 (0.16)	1.92 (0.39)	0.68 (0.23)
Investimento	12.00 (4.05)	0.02 (0.01)	0.83 (0.21)	0.87 (0.20)	1.56 (0.58)	0.67 (0.22)
Malattia	11.00 (3.28)	0.02 (0.01)	0.74 (0.17)	0.76 (0.17)	1.91 (0.41)	0.63 (0.22)
Previdenza	11.00 (3.32)	0.02 (0.01)	0.75 (0.17)	0.78 (0.17)	1.90 (0.43)	0.62 (0.21)
Protezione	10.00 (3.15)	0.02 (0.01)	0.73 (0.16)	0.73 (0.16)	1.91 (0.41)	0.58 (0.21)
Rischi Civili	9.00 (2.92)	0.02 (0.01)	0.71 (0.16)	0.70 (0.14)	1.93 (0.38)	0.55 (0.20)
Risparmio	12.00 (3.82)	0.03 (0.02)	0.83 (0.20)	0.88 (0.19)	1.69 (0.54)	0.64 (0.22)
Vita Collettive	12.00 (3.11)	0.07 (0.19)	0.87 (0.20)	0.93 (0.12)	1.55 (0.57)	0.63 (0.22)

Tabella 2.2: Valori medi delle misure di rete locali relative ai prodotti delle reti di secondo livello, e relativa deviazione standard (tra parentesi)

- Transitività globale: questa misura viene anche detta *coefficiente di clustering*, e corrisponde alla proporzione di triangoli presenti nella rete, rispetto al numero di possibili triangoli.

In Tabella 2.3 si riporta la distribuzione delle statistiche globali relative alle reti di prodotto associate alle agenzie. Come si è notato dai valori delle misure locali, le reti risultano complessivamente avere un buon numero connessioni e conformazioni di rete (la densità e la transitività, sono pressoché le stesse tra le varie reti), e si nota in aggiunta, la presenza di qualche rete senza nessuna connessione. Si tratta di quelle agenzie di piccole dimensioni che si sono notate nelle analisi descrittive: significa quindi che non vi è, per quelle agenzie, un numero sufficiente di clienti multi-prodotto in grado di definire delle connessioni.

	Densità	Diametro	Raggio	Transitività
Min.	0.00	0.00	0.00	0.00
I quartile	0.74	2.00	1.00	0.68
Mediana	0.74	2.00	1.00	0.70
Media	0.67	1.93	1.23	0.62
III quartile	0.74	2.00	2.00	0.71
Max.	0.74	2.00	2.00	0.84

Tabella 2.3: Distribuzione delle misure di rete globali relative alle 135 reti di prodotti

Data una descrizione delle reti in termini di misure, è di interesse verificare quanto le reti di secondo livello siano informative sulle connessioni presenti nella rete di primo livello, ovvero se, a parità di strutture di dipendenze delle reti di prodotto, le agenzie risultano connesse, e quindi simili, all'interno della propria rete. A tale scopo si considera nel seguente paragrafo un modello base per dati di rete, e lo si applica ai dati di rete della struttura di vendita.

2.3 I modelli classici di rete

I modelli ERGM (*Exponential Random Graph Models*) rappresentano la famiglia classica di modelli statistici per dati di rete, e sono pensati in analogia ai modelli lineari generalizzati (GLM). Infatti, sono formulati in modo da rendere possibile l'estensione dei principi di costruzione, stima e confronto tra modelli della statistica classica. Tuttavia la complessità delle strutture di dipendenza tra i legami relazionali, comportano una serie di adattamenti ed estensioni del modello non banali. Per questo motivo i modelli ERGM hanno un buon potenziale a livello teorico, ma nella pratica spesso si adattano in maniera povera ai dati.

Si consideri il grafo relativo alla rete di primo livello $\mathcal{G}$ definito sull'insieme di nodi $\mathcal{N}$ con associata matrice di adiacenza $\boldsymbol{X}$. Un modello ERGM specifica una forma di tipo esponenziale per la distribuzione congiunta degli elementi di X

$$p(\boldsymbol{X} = \boldsymbol{x}; \theta) = \frac{1}{\kappa(\boldsymbol{\theta})} \exp(\boldsymbol{\theta}^T g(\boldsymbol{x})) = \exp(\boldsymbol{\theta}^T g(\boldsymbol{x}) - log(\kappa(\boldsymbol{\theta}))) \tag{2.5}$$

in cui

- $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ è un vettore di p parametri
- $g(\boldsymbol{x})$ è un vettore di p statistiche di rete $\kappa(\boldsymbol{\theta})$
- $\kappa(\boldsymbol{\theta})$ è una costante di normalizzazione definita come $\kappa(\boldsymbol{\theta}) = \sum_{z \in \mathcal{X}} \exp\{\kappa(\boldsymbol{\theta}^T g(Z)\}$, che non dipende da X ma che risulta in genere difficile da calcolare.

Si noti che l'espressione in (2.5) appartiene ad una famiglia esponenziale; di conseguenza il nome *Exponential random graph model*. I modelli classici sono soliti considerare gli archi X_{ij} siano variabili casuali indipendenti tra loro; ad ogni statistica di rete che comprende tre o più vertici è quindi associato un parametro $\theta = 0$. Si ha in generale un modello nella forma

$$p(\boldsymbol{X} = \boldsymbol{x}; \kappa(\boldsymbol{\theta})) = \prod_{i<j} p(\boldsymbol{X}_{ij} = x_{ij}) = \prod_{i<j} \pi_{ij}^{(x_{ij})} (1 - \pi_{ij})^{(1-x_{ij})} = \frac{1}{\kappa(\boldsymbol{\theta})} \exp\big(\sum_{i<j} \boldsymbol{\theta}_{ij} x_{ij}\big). \tag{2.6}$$

dove $logit(\pi_{ij}) = \theta_{ij}$. Dato che i parametri sono tanti quanti le variabili casuali osservate, è necessario definire una qualche forma funzionale per poterne ottenere la stima. Il modello di Erdős & Rényi (1959) considera un assunzione di omogeneità delle connessioni all'interno della rete, fissando $\theta_{ij} = \boldsymbol{\theta}$ per ogni $i, j \in \mathcal{N}$. Si tratta di un modello semplice da stimare, ma del tutto non realistico, in quanto non è in grado di cogliere eventuali effetti di eterogeneità degli attori, omofilia, o transitività.

Una classe più flessibile di modelli ERGM è quella costituita dai modelli markoviani, basati appunto sul concetto di *dipendenza markoviana* introdotto da Frank & Strauss (1986): due nodi sono dipendenti se, condizionatamente al resto della rete, hanno almeno un nodo in comune. L'idea alla base è che la probabilità, in termini di *log-odds*, di osservare una certa configurazione di rete del grafo $\mathcal{G}$ è proporzionale al numero di archi presenti nella rete; è di comune pratica introdurre nel modello statistiche di ordine maggiore sull'intera struttura di

rete, come ad esempio i conteggi dei triangoli $T(X)$ o delle configurazioni k-stelle $S_k(X)$[2], ottenendo il seguente modello:

$$p(\boldsymbol{X} = x; \boldsymbol{\theta}) = \frac{1}{\kappa(\boldsymbol{\theta})} \exp\Big\{ \sum_{k=1}^{N-1} \theta_k S_k(X) + \theta_T T(X) \Big\}. \tag{2.7}$$

Nella pratica, il modello markoviano spesso produce delle stime che si adattano in maniera povera ai dati, e sarebbe necessaria l'introduzione di statistiche di ordine maggiore, che rendono tuttavia problematica la stima stessa del modello; ulteriori statistiche di rete sono state proposte negli anni come soluzione, si rimanda a Snijders et al. (2006) per una discussione.

I modelli descritti si basano esclusivamente su informazioni di tipo endogeno, ovvero riguardanti la rete stessa; è plausibile però che la probabilità di un arco tra due nodi non dipenda solamente dallo stato delle altre coppie di vertici, ma anche dagli attributi associati ai vertici stessi. É possibile includere covariate nel modello in forma di statistica di rete all'interno del termine esponenziale in (2.5) come

$$g(x, z) = \sum_{1 \le i \le j \le N} x_{ij} h(\boldsymbol{z}_i, \boldsymbol{z}_j) \tag{2.8}$$

dove $h(\cdot)$ è una funzione di tipo simmetrico che misura la similarità dei vettori di attributi $\boldsymbol{z}_i$ e $\boldsymbol{z}_j$ osservati per i nodi i e j. Comunemente si considera per h funzioni del tipo $h(\boldsymbol{z}_i, \boldsymbol{z}_j) = \boldsymbol{z}_i + \boldsymbol{z}_j$ o $h(\boldsymbol{z}_i, \boldsymbol{z}_j) = \mathbb{I}(\boldsymbol{z}_i = \boldsymbol{z}_j)$, che specificano rispettivamente un effetto principale e di secondo ordine (*omofilia*).

Considerano i dati, si sono stimati due modelli di tipo ERGM per la rete di primo livello, inserendo le misure delle reti di secondo livello come covariate; il primo considerando per ogni rete le misure locali per ogni nodo-prodotto (numerati da 1 a 15), mentre nel secondo sono state riassunte come statistiche di rete di tipo globale. Le variabili di tipo discreto sono state inserite nel modello specificando un effetto di omofilia (*nodematch*), mentre quelle di tipo continuo sono trattate come effetti principali (*nodecov*); si sono escluse le misure di eccentricità locali, in quanto direttamente correlate alle misure di raggio e diametro. In Tabella 2.4 si riportano i coefficienti stimati.

Il modello con covariate le misure locali pare adattarsi leggermente meglio ai dati in analisi, anche se risulta comunque povero: la pratica comune per determinare la bontà del modello consiste nel simulare un certo numero di reti dal modello stimato e confrontare le distribuzioni delle misure di rete dei grafi simulati, con quella osservate originariamente nei dati; in Figura 2.1 si riportano alcuni grafici di diagnostica, dai quali si può notare che il modello non pare adattarsi perfettamente bene ai dati.

[2]Una stella è un grafo in cui un nodo, detto centro, è adiacente a tutti gli altri e questi sono adiacenti solo al centro. Una stella con k+ 1 nodi viene normalmente indicata con S_k

	Modello con misure locali	Modello con misure globali
archi	29.15 (3.87)***	−5.29 (1.01)***
nodematch.diametro	0.63 (0.38)	1.68 (0.34)***
nodematch.raggio	−0.13 (0.06)*	0.04 (0.05)
nodematch.grado.1	0.67 (0.07)***	
nodematch.grado.2	0.01 (0.07)	
nodematch.grado.3	0.10 (0.07)	
nodematch.grado.4	0.19 (0.07)**	
nodematch.grado.5	0.20 (0.07)**	
nodematch.grado.6	0.10 (0.07)	
nodematch.grado.7	−0.02 (0.07)	
nodematch.grado.8	0.11 (0.06)	
nodematch.grado.9	0.10 (0.07)	
nodematch.grado.10	0.09 (0.08)	
nodematch.grado.11	0.56 (0.09)***	
nodematch.grado.12	−0.06 (0.06)	
nodematch.grado.13	−0.02 (0.08)	
nodematch.grado.14	0.12 (0.08)	
nodematch.grado.15	0.03 (0.08)	
nodecov.densità	131.31 (12.60)***	−14.06 (2.00)***
nodecov.transitività	−6.48 (4.34)	1.38 (0.73)
nodecov.betweenness.1	47.96 (6.60)***	
nodecov.betweenness.2	38.77 (18.03)*	
nodecov.betweenness.3	7.57 (26.79)	
nodecov.betweenness.4	71.93 (24.03)**	
nodecov.betweenness.5	138.36 (19.93)***	
nodecov.betweenness.6	178.38 (33.88)***	
nodecov.betweenness.7	6.33 (32.25)	
nodecov.betweenness.8	−20.96 (32.24)	
nodecov.betweenness.9	−51.75 (18.51)**	
nodecov.betweenness.10	69.49 (18.94)***	
nodecov.betweenness.11	121.48 (29.23)***	
nodecov.betweenness.12	89.49 (39.10)*	
nodecov.betweenness.13	−184.75 (28.30)***	
nodecov.betweenness.14	56.62 (6.17)***	
nodecov.betweenness.15	48.37 (6.86)***	
nodecov.closeness.1	−27.59 (3.59)***	
nodecov.closeness.2	−1.45 (4.92)	
nodecov.closeness.3	21.73 (6.37)***	
nodecov.closeness.4	−21.77 (4.68)***	
nodecov.closeness.5	−39.80 (5.61)***	
nodecov.closeness.6	−22.83 (6.71)***	
nodecov.closeness.7	−25.28 (5.44)***	
nodecov.closeness.8	−2.15 (7.64)	
nodecov.closeness.9	2.15 (4.36)	
nodecov.closeness.10	−24.37 (8.74)**	
nodecov.closeness.11	−22.74 (6.29)***	
nodecov.closeness.12	−19.88 (9.11)*	
nodecov.closeness.13	66.30 (9.38)***	
nodecov.closeness.14	−27.38 (4.23)***	
nodecov.closeness.15	1.20 (4.05)	
nodecov.autovalori.1	13.17 (2.85)***	
nodecov.autovalori.2	−6.72 (2.81)*	
nodecov.autovalori.3	−18.82 (3.20)***	
nodecov.autovalori.4	7.18 (2.39)**	
nodecov.autovalori.5	12.78 (2.97)***	
nodecov.autovalori.6	0.91 (2.75)	
nodecov.autovalori.7	6.55 (2.58)*	
nodecov.autovalori.8	−3.04 (3.86)	
nodecov.autovalori.9	−3.02 (2.82)	
nodecov.autovalori.10	4.02 (4.21)	
nodecov.autovalori.11	7.91 (2.84)**	
nodecov.autovalori.12	2.82 (4.32)	
nodecov.autovalori.13	−32.30 (4.73)***	
nodecov.autovalori.14	12.36 (3.16)***	
nodecov.autovalori.15	−9.72 (3.55)**	
nodecov.betweenness		−110.15 (20.02)***
nodecov.closeness		8.17 (1.72)***
nodecov.grado		0.13 (0.03)***
nodecov.eccentricità		2.33 (0.32)***
nodecov.autovalori		1.36 (0.54)*
AIC	10253.39	11499.92
BIC	10715.54	11571.02
Log Likelihood	−5061.70	−5739.96

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Tabella 2.4: Modelli ERGM per la rete di primo livello, con misure delle reti di secondo livello come covariate.

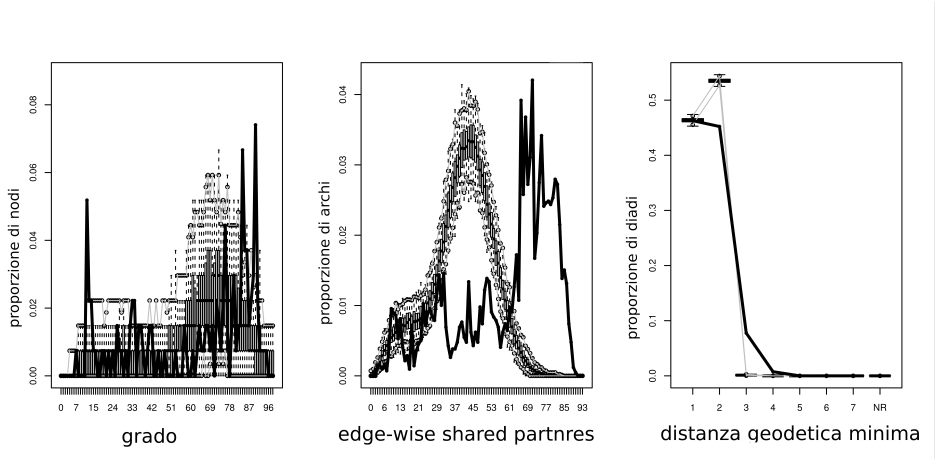

Figura 2.1: Grafici di diagnostica sulla bontà di adattamento del modello ERGM ai dati, rispetto al grado, il raggio ed il numero di vicini condivisi dalle coppie di nodi che sono collegate

Un modello ERGM fornisce un'interpretazione per i coefficienti dei parametri stimati analoga a quella dei modelli GLM, ma non sono garantite a livello teoriche le stesse proprietà asintotiche nel calcolo dei valori-p, che vengono considerati come statistiche aggiuntive. Allo scopo di interpretare i coefficienti, si consideri la probabilità di avere una connessione tra due vertici, condizionata allo stato degli archi nel resto della rete e si indichi con $\boldsymbol{X}_{(-ij)}$ la matrice di adiacenza $\boldsymbol{X}$ a meno dell'elemento X_{ij}; la distribuzione dell'elemento X_{ij} condizionata a $\boldsymbol{X}_{(-ij)}$ è di tipo bernoulliano e soddisfa l'espressione

$$\log\Big[\frac{p(X_{ij}=1|\boldsymbol{X}_{(-ij)}=\boldsymbol{x_{(-ij)}};\boldsymbol{\theta})}{p(X_{ij}=0|\boldsymbol{X}_{(-ij)}=\boldsymbol{x_{(-ij)}};\boldsymbol{\theta})}\Big]=\boldsymbol{\theta}^T\Delta_{ij}(\boldsymbol{x}) \tag{2.9}$$

dove $\Delta_{ij}(\boldsymbol{x})$ è una *statistica di cambio*, che descrive la differenza tra il vettore di statistiche $g(\boldsymbol{x})$, quando y_{ij} è pari a 1 o 0.

I coefficienti stimati dal modello ERGM, possono quindi essere interpretati come in termini di *odds-ratio* condizionato delle probabilità di connessione. Ad esempio, la corrispondenza del grado del nodo-prodotto Previdenza (*nodematch.grado.11*) nelle reti di secondo livello di due agenzie, la probabilità di connessione tra le due, aumenta l'*odds* di un fattore pari a $\exp(0.56) = 1.75$, a parità delle altre statistiche. Si hanno maggiori informazioni in merito alle strutture di dipendenza delle reti di prodotto, ma risulta ancora complicato definire una campagna di marketing sulla base del modello principalmente per due motivi:

1. il confronto tra agenzie avviene a coppie, ed è complicato determinarne un gruppo sul quale agire secondo una strategia comune, senza l'uso di informazioni esogene; ad esempio informazioni che permettono di fissare alcune agenzie come "ideali", e operare in modo che le altre agenzie risultino simili a queste, ovvero aumentare la probabilità di connessione tra dei nodi specifici, agendo sulle covariate.

2. agire sulle covariate all'interno della rete di prodotto è comunque complicato, in quanto non si ha un'indicazione precisa su quali prodotti agire per aumentare, ad esempio, la centralità di un nodo.

Sulla base di tali motivazioni, è quindi necessario definire un modello che ci permetta di ottenere (i) una rappresentazione ridotta delle agenzie in gruppi e (ii) una caratterizzazione di tali gruppi in termini di tipologia di politiche di vendita (rete di prodotti). Come soluzione al primo punto, è possibile considerare un'altra classe di modelli di rete, che tiene conti di raggruppamenti di nodi all'interno della rete che condividono una stessa struttura di connessioni, i modelli a blocchi stocastici.

2.4 Modelli a blocchi stocastici

L'obiettivo di questo tipo di modelli è quello di partizionare l'insieme dei vertici in sottoinsieme chiamati *blocchi* in modo tale che la composizione dei blocchi e la struttura delle connessioni tra di essi, sia in grado di rappresentare le caratteristiche relazionali principali del grafo. L'assunzione alla base è quella di *equivalenza strutturale* (Lorrain & White, 1971): attori caratterizzati dagli stessi attributi, e quindi aventi le stesse connessioni con gli altri nodi sono strutturalmente equivalenti (appartengono alla stessa partizione). Fienberg & Wasserman (1981) e Holland et al. (1983) generalizzano il concetto in termini stocastici: un modello a blocchi stocastici può essere definito come una distribuzione di probabilità (o una famiglia di distribuzioni) su grafi in cui l'insieme di vertici è partizionato in sottoinsieme chiamati blocchi, tali che la distribuzione di probabilità del grafo è invariante rispetto a permutazione dei vertici all'interno dei blocchi. Sotto tale modello, la probabilità di connessione tra due nodi dipende solamente dai relativi blocchi di appartenenza, chiamati anche *colori*. Nelle applicazioni pratiche solitamente non si è a conoscenza dei blocchi a priori, e si applica quindi un approccio simile a quello dei modello di mistura, in termini di *mistura di modelli ERGM* (Daudin et al., 2008). Il modello a blocchi stocastici è formalizzato in Nowicki & Snijders (2001), al quale si fa riferimento nella seguente definizione.

Sia W la matrice di adiacenza associata al grafo $\mathcal{G}$ sull'insieme di N nodi, i quali appartengono a B differenti categorie, definite come *blocchi* o *colori*. Siano i blocchi rappresentati da una variabile casuale $\boldsymbol{y} = (y_1, \ldots, y_N)$ associata ai vertici, con valori in $0, \ldots, B$, tale che $y_i = k$ se il vertice i ha colore k, per $k = 1, \ldots, B$ e $i \in \{1, \ldots, N\}$. Dato un vertice i di colore k ed uno j di colore l, la probabilità di connessione tra i due vertici può essere scritta come $p(w_{ij}|y_i = k, y_j = l) = \eta_{kl}$, in cui η_{kl} è la probabilità di connessione tra due nodi specifica della classe. Quindi, data la colorazione di due nodi, si può definire la presenza di un arco tra i due, come un'estrazione indipendente da una variabile casuale Bernoulliana con probabilità di successo dipendente dai colori dei blocchi; si ottiene quindi una rappresentazione ridotta della matrice di adiacenza, definita in termini probabilistici, chiamata *immagine*

$$\begin{pmatrix} \eta_{11} & \eta_{12} & \cdots & \eta_{1B} \\ \eta_{21} & \eta_{22} & \cdots & \eta_{2B} \\ \vdots & \vdots & \ddots & \vdots \\ \eta_{B1} & \eta_{B2} & \cdots & \eta_{BB} \end{pmatrix} \tag{2.10}$$

Il modello di Nowicki & Snijders (2001) prevede che i nodi siano partizionati in blocchi

solo sulla base delle strutture interne alla rete; Tallberg (2004) ne propone un'estensione allo scopo di introdurre delle covariate per la stima dei blocchi: le probabilità di appartenenza al blocco condizionate agli attributi dei nodi secondo un modello multinomiale di tipo probit. Si potrebbe utilizzare tale approccio per i dati a disposizione, in modo da ottenere una rappresentazione in gruppi delle agenzie, in cui ogni gruppo è caratterizzato da reti di prodotto con specifiche strutture di dipendenza. Tuttavia si è concluso nell'analisi dei risultati ottenuti tramite modello ERGM, che una rappresentazione delle reti di prodotto in termini di misure di rete, aiuta sì la comprensione delle strutture di dipendenza, ma non permette di avere una visione abbastanza approfondita delle connessioni tra prodotti.

Capitolo 3

Un modello bayesiano per una rete di reti

La maggior parte dei metodi statistici tipici dell'analisi delle reti sociali sono concepiti come modelli per una sola rete, mentre i dati disponibili sono un caso in cui, fissati i nodi di una rete, si osserva una popolazione di insiemi di archi. Infatti le reti di prodotti condividono gli stessi indentici nodi, e sono connessi tra loro in maniera differente a seconda dell'agenzia alla quale sono associate. In aggiunta, ogni agenzia è un nodo all'interno di una rete che descrive una relazione di similarità tra esse. Siamo quindi alla ricerca di un modello statistico che sia in grado di (i) fornire una rappresentazione ridotta delle agenzie in gruppi e (ii) caratterizzare tali gruppi in termini di tipologia di politiche di vendita, attraverso le reti di prodotto.

Allo scopo di fornire una rappresentazione della struttura complessa dei dati, si considera in questo capitolo un contesto bayesiano di tipo non parametrico. Si propone un modello a blocchi stocastici per la rete di primo livello, in grado di sfruttare appieno l'informazione delle reti di secondo livello. Osservazioni di reti multiple, come le nostre reti di prodotti, sono disponibili in diversi ambiti di ricerca (quali ad esempio le neuroscienze o la biologia), ma i metodi di analisi in letteratura, sono soliti studiare solamente le strutture di dipendenza comuni alla popolazione di rete, o ridurre il campione ad un insieme di statistiche, come è stato proposto nel Capitolo 2. Un approccio in grado di rappresentare in termini statistici la distribuzione di una popolazione di osservazioni di rete, è stato di recente proposto da Durante et al. (2015), che descrivono un modello bayesiano non parametrico in grado di fornire una rappresentazione sia delle strutture comuni alle reti, sia le caratteristiche peculiari di classi di reti.

Nel seguente paragrafo si riporta il modello proposto da Durante et al. (2015), utilizzato per ottenere una rappresentazione ridotta delle reti di prodotti attraverso una procedura di *clustering*; si presenta nel paragrafo successivo, il modello a blocchi stocastici da noi proposto per la definizione congiunta della reti di agenzie e le rispettive reti di prodotti.

3.1 Modello di clustering per le reti di secondo livello

Siano $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_N$ osservazioni multiple relative ad grafo semplice [1] $\mathcal{H}$ definito su un'insieme di nodi $\mathcal{V}$ di cardinalità $|\mathcal{V}| = V$, come le nostre reti di prodotti. Ogni osservazione $\boldsymbol{A}_i$ è una matrice di adiacenza di dimensione $V \times V$ con elementi $A_{vu,i} = A_{uv,i} \in \{0, 1\}$ codificanti la presenza o meno di una conessione tra i nodi v e u per l'osservazione i. Dato che è di interesse un grafo semplice, la matrice di adiacenza $\boldsymbol{A}_i$ è simmetrica, si può considerare solamente la rispettiva matrice triangolare inferiore, che denotiamo con il vettore $\mathcal{L}(\boldsymbol{A}_i) = (A_{21,i}, A_{31,i}, \ldots, A_{V1,i}, A_{32,i}, \ldots, A_{V2,i}, \ldots, A_{V(V-1),i})^T \in \mathcal{Y}_V = \{0, 1\}^{V(V-1)/2}$.

I vettori $\mathcal{L}(\boldsymbol{A}_1), \ldots, \mathcal{L}(\boldsymbol{A}_n)$ sono realizzazioni di una variabile aleatoria di Bernoulli multivariata $\mathcal{L}(\boldsymbol{\mathcal{A}})$, le cui componenti sono delle variabili bernoulliane, con associata distribuzione di probabilità $p_{\mathcal{L}(\boldsymbol{A})}$. [2] Dato che il numero di possibili configurazioni di rete su un'insieme di nodi V è finito, si può pensare $\mathcal{L}(\boldsymbol{A})$ come una variabile aleatoria discreta il cui supporto è dato dai vettori delle possibili configurazione di rete $\boldsymbol{a} \in \mathcal{Y}_V$; ad esempio, la variabile casuale $\mathcal{L}(\boldsymbol{A})$ associata ad un grafo con $V = 3$ nodi, avrà $2^{V(V-1)/2)} = 8$ possibili configurazioni di rete $\{(0, 0, 0); (1, 0, 0); \ldots; (1, 1, 1)\}$. Sotto la condizione $\sum_{\boldsymbol{a} \in \mathcal{Y}_V} p_{\mathcal{L}(\boldsymbol{A})}(\boldsymbol{a}) = 1$, sono necessari $2^{V(V-1)/2}$ parametri per caratterizzare la distribuzione di probabilità $p_{\mathcal{L}(\boldsymbol{A})}(\boldsymbol{a}) = Pr\{\mathcal{L}(\boldsymbol{A}) = \boldsymbol{a}\}, \boldsymbol{a} \in \mathcal{Y}_V$; tuttavia, il numero di parametri cresce velocemente all'aumentare delle osservazioni, anche in contesti in cui il numero di nodi V è relativamente piccolo, tanto che risulta impossibile ottenere una stima della distribuzione di probabilità $p_{\mathcal{L}(\boldsymbol{A})}(\boldsymbol{a})$ con metodi non parametrici, senza imporre dei vincoli.

Durante et al. (2015) propongono un nuovo approccio per la stima di tale distribuzione di probabilità, basato su un modello bayesiano di tipo non paramtrico; si tratta del primo metodo trovato in letteratura in grado di definire modello generativo probabilistico per descrivere la distribuzione di una popolazione di osservazioni multiple di rete. In particolare, viene asssociato un modello di mistura alla distribuzione di probabilità $p_{\mathcal{L}(\boldsymbol{A})}(\boldsymbol{a})$, in cui gli individui sono allocati in classi in basa alla relativa struttura di rete. All'interno di ogni classe, le probabilità di connessione tra i nodi sono messe in relazione a delle misure latenti di similarità espresse in termini matriciali, secondo un legame di tipo logistico; le matrici similarità sono a loro volta scomposte come somma di una componente comune a tutta la popolazione di reti e una componente specifica della classe. Quest'ultima componente è definita su uno spazio latente di basso rango associati all'insiemi dei nodi $\mathcal{V}$ e tiene conto della struttura specifica della classe di rete. La distribuzione di probabilità $p_{\mathcal{L}(\boldsymbol{A})}(\boldsymbol{a})$, è quindi rappresentata come una mistura di fattorizzazioni a basso rango che riduce la dimensione del

[1] Richiamando la definizione enunciata al Paragrafo 1.2: un grafo semplice è un grafo che non contiene né cappi (*self-loops*) né archi multipli, di tipo non orientato e non pesato.

[2] Una distribuzione bernoulliana multivariata, è una distribuzione di probabilità usualmente associata ai grafi di tipi indiretto, atta a rappresentare la distribuzione di probabilità del grafo stesso; le componenti di tale variabile sono sì delle bernoulliana, ma presentano un struttura di dipendenza. A titolo esplicativo si consideri il caso di un vettore casuale bivariato di Bernoulli (Y_1, Y_2) che assume valori $\{(0, 0), (0, 1), (1, 0), (1, 1)\}$. Sia $p_{ih} = p(Y_i = i, Y_j = j), i, j = 0, 1$. La distribuzione di probabilità della variabile è data da

$$p(\boldsymbol{Y} = \boldsymbol{y}) = p(y_1, y_2) = p_{11}^{y_1 y_2} p_{10}^{y_1(1-y_2)} p_{01}^{y_2(1-y_1)} p_{00}^{(1-y_1)(1-y_2)} \tag{3.1}$$

Analogamente sono definite distribuzioni di ordine superiore. Si rimanda a Dai et al. (2013) per le proprietà e risultati al riguardo.

campione maniera automatica, attraverso una procedura di *clustering* indotta dal modello stesso.

Definita la classe di appartenenza di una generica osservazione di rete i, è possibile considerare gli archi $\mathcal{L}(\mathcal{A}_i)_l, l = 1, \ldots, V(V-1)/2$ come variabili aleatorie bernoulliane condizionatamente indipendenti, data la corrispondente probabilità di connessione $\pi_{il}^{(h)} = Pr\{\mathcal{L}(\mathcal{A}_i)_l = 1 | G_i = h\}, l = 1, \ldots, V(V-1)/2, h = 1, \ldots, H$, e ottenere per ogni classe la corrispondente distribuzione di probabilità:

$$p_{\mathcal{L}(\boldsymbol{A}_i)}(\boldsymbol{a}|G_i = h) = \prod_{l=1}^{V(V-1)/2} (\pi_l^{(h)})^{a_l}(1-\pi_l^{(h)})^{1-a_l}. \tag{3.2}$$

I vettori di probabilità di connessione $\boldsymbol{\pi_i}, i = 1, \ldots, N$ sono considerati come effetti casuali associati ad una comune misura di probabilità discreta P definita sulle classi latenti. In particolare si ha

$$\begin{aligned} \mathcal{L}(\boldsymbol{A}_i)_l| &\sim Bern(\pi_{il}), \quad l = 1, \ldots, V(V-1)/2, \quad i = 1, \ldots, N \\ \boldsymbol{\pi}_i &\sim P = \sum_{h=1}^{H} \nu_h \delta_{\boldsymbol{\pi}^{(h)}}, \quad \boldsymbol{a} \in \mathcal{Y}_V \end{aligned} \tag{3.3}$$

in cui $\delta_{\boldsymbol{\pi}}^{(h)}$ rappresenta una massa di probabilità (atomo) concentrata in $\boldsymbol{\pi}^{(h)}$ e ν_h la probabilità che una rete casualmente estratta, sia allocata nella classe h. La scelta di tale misura permette al modello di indurre una procedura di *clustering* sulla popolazione di reti in H classi latenti, in cui le reti appartenenti alla stessa classe hanno associato uno stesso vettori di probabilità di connessione $\boldsymbol{\pi}^{(h)}$.

Un'ulteriore riduzione della dimensionalità del problema, viene ottenuta attraverso una scomposizione in fattori del vettore di probabilità delle connessioni $\boldsymbol{\pi}^{(h)}$., in grado di tenere conto dell'intera struttura di rete del campione e allo stesso tempo fornire una rappresentazione a basso rango di $\boldsymbol{\pi}^{(h)}$, la cui dimensione può variare a seconda della classe, come:

$$\boldsymbol{\pi}^{(h)} = \{1 + \exp[-\boldsymbol{Z} - \boldsymbol{D}^{(h)}]\}^{-1}, \quad \boldsymbol{D}^{(h)} = \mathcal{L}(\boldsymbol{X}^{(h)}\boldsymbol{\Lambda}^{(h)}\boldsymbol{X}^{(h)T}) \tag{3.4}$$

in cui la funzione logistica è applicata elemento per elemento. L'equazione (3.4) definisce il vettore dei *log-odds* delle probabilità di connessione $\boldsymbol{S}^{(h)} = (S_1^{(h)}, \ldots, S_{V(V-1)/2}^{(h)})$, come la somma di un vettore di similarità $\boldsymbol{Z} \in \mathbb{R}^{V(V-1)/2}$ comune a tutte le reti, e di un vettore di variabilità classe-specifico $\boldsymbol{D}^{(h)} \in \mathbb{R}^{V(V-1)/2}$. Mentre il vettore delle similarità $\boldsymbol{Z}$ racchiude in sé la porzione di informazione comune a tutte le reti e non ha una struttura definita, il vettore $\boldsymbol{D}^{(h)}$ è specifico della classe ed è definito come $\boldsymbol{D}^{(h)} = \mathcal{L}(\boldsymbol{X}^{(h)}\boldsymbol{\Lambda}^{(h)}\boldsymbol{X}^{(h)T})$. L'elemento $\boldsymbol{X}^{(h)} \in \mathbb{R}^{V \times R}$ è una matrice di R coordinate latenti dell'insieme di nodi $\mathcal{V}$ su uno spazio latente di dimensione R (usualmente $R \ll V$), mentre $\boldsymbol{\Lambda}^{(h)}$ è una matrice diagonale di elementi $(\lambda_1^{(h)}, \ldots, \lambda_R^{(h)})^T = \boldsymbol{\lambda}^{(h)} \in \mathbb{R}_{\geq 0}^R$. Si denota con $\mathbb{R}_{\geq 0}^R$ lo spazio dei vettori di R elementi non negativi. La scomposizione in fattori separata per ogni $\boldsymbol{D}^{(h)}$ permette una

rappresentazione altamente flessibile della struttura di dipendenze all'interno della classe h in quanto la dimensione dello spazio latente può variare all'interno di ogni classe.

Marginalizzando la distribuzione congiunta delle osservazioni $\mathcal{L}(\mathcal{A}_i)$ rispetto ai corrispondenti vettori di probabilità di connessione π_i, si ottiene la seguente rappresentazione della distribuzione di probabilità $p_{\mathcal{L}(\boldsymbol{A})}$ associata alla variabile casuale $\mathcal{L}(\boldsymbol{A})$:

$$p_{\mathcal{L}(\boldsymbol{A})}(\boldsymbol{a}; \nu_1, \ldots, \nu_H; \boldsymbol{\pi}^{(1)}, \ldots, \boldsymbol{\pi}^{(H)}) = \sum_{h=1}^{H} \nu_h \prod_{l=1}^{V(V-1)/2} \pi_l^{a_l} (1-\pi)^{1-a_l} \tag{3.5}$$

per ogni configurazione $\boldsymbol{a} \in \mathcal{Y}_V$, in cui ogni $\boldsymbol{\pi}^{(h)}$ è fattorizzato come in (3.4), per ogni $h = 1, \ldots, H$.

3.1.1 Distribuzione a priori

Durante et al. (2015) definiscono delle condizioni generali sulle distribuzioni a priori delle quantità $\boldsymbol{Z} \sim \Pi_Z$, $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_H)^T \sim \Pi_\nu$, $\boldsymbol{X}^{(h)} \sim \Pi_X$ and $\boldsymbol{\lambda_{(h)}} \sim \Pi_\lambda$ per garantire che il supporto della distribuzione a priori includa tutte le possibili configurazioni di rete. Siano H e R i limiti superiori, rispettivamente per il numero di reti latenti H^0 ed il numero di coordinate latenti R^0. Le distribuzioni a priori sono definite con lo scopo di favorire l'eliminazione delle dimensioni ridondanti, in modo che la distribuzione a posteriori si concentri in $\nu_h \approx 0$ per $h > H^0$ e $\lambda_r \approx 0$ per $r > R^{0(h)}$ con $R^{0(h)}$ rapprensentante il numero sufficiente di coordinate richiesto per rappresentare il vero vettore di probabilità $\boldsymbol{\pi}^{0(h)}$ nella fattorizzazione a rango ridotto di ogni $h = 1, \ldots, H$. A tale scopo è definita una distribuzione a priori con due livelli di penalizzazione:

1. Il primo livello riduce la dimensione del campione di osservazioni di rete, attraverso una procedura di *clustering* delle osservazioni. Dato che non vuole si imporre a priori il numero di classi latenti, ma lasciare che siano i dati a determinarlo, è pratica comune nei modelli bayesiani non parametrici, definire la misura di probabilità P in (3.3) come un *processo di Dirichlet* (Ferguson, 1973). Un processo di Dirichlet è una generalizzazione infinito-dimensionale della distribuzione di Dirichlet, comunemente utilizzata come coniugata a priori di distribuzioni discrete non parametriche. Una formulazione di tipo costruttivo del processo, è data dalla rappresentazione *stick-breaking* (Sethuraman, 1994)

$$\boldsymbol{\pi}_i \sim P = \sum_{h=1}^{\infty} \nu_h \delta_{\boldsymbol{\pi}}^{(h)}, \quad \boldsymbol{a} \in \mathcal{Y}_V \tag{3.6}$$

$$\nu_h = V_h \prod_{j<h} (1 - V_j), \quad V_h \sim Beta(1, \alpha) \tag{3.7}$$

in cui $\delta_{\boldsymbol{\pi}}^{(h)}$ rappresenta una massa di probabilità (atomo) concentrata in $\boldsymbol{\pi}^{(h)}$ e ν_h la probabilità che una rete casualmente estratta, sia allocata nella classe h. La dicitura *stick-breaking* deriva dalla metafora associata alla costruzione dei pesi ν_h. Si considera

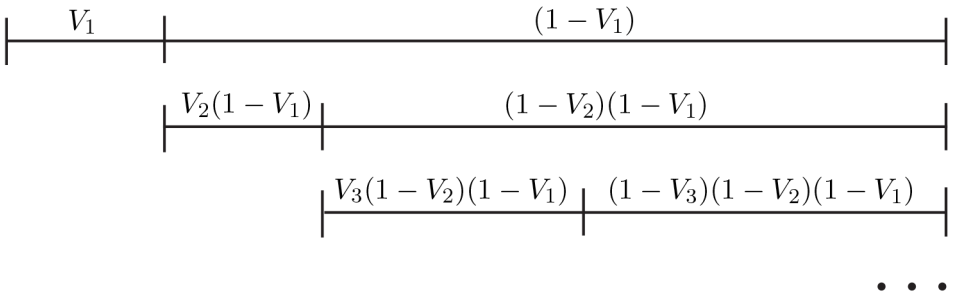

Figura 3.1: Rappresentazione grafica del processo *stick-breaking*

un'asta di lunghezza unitaria, che viene spezzata in un punto aleatorio V_1; la lunghezza del segmento ottenuto viene assegnata a ν_1, e allo stesso modo, il processo viene iterato per ottenere gli ulteriori pesi $\nu_2, \nu_3, \ldots$ (Figura 1). Si noti che la somma dei pesi è per costruzione pari a uno. L'iperparametro α regola il numero di atomi del processo ed è detto *parametro di dispersione*: per $\alpha \to 0$ le realizzazioni del processo sono concentrate attorno un singolo atomo, mentre per $\alpha \to \infty$ le realizzazioni si approssimano al continuo.

Tale rappresentazione permette che le osservazioni di rete vengano allocate in H classi, e che alle reti all'interno della stessa classe h corrisponda il medesimo vettore di probabilità delle connessioni $\boldsymbol{\pi}^{(h)}$. A livello teorico, il numero di classi latenti H nella popolazione è potenzialmente infinito, ma Durante et al. (2015) utilizzano i risultati in Ishwaran & Zarepour (2002) che dimostrano non solo che il numero di classi nella popolazione è quasi certamente finito, ma che il processo di Dirichlet è aprossimabile all'omonima distribuzione.

Quindi, fissato un limite superiore per le classi latenti, la distribuzione a priori per i pesi sulle classi è definita come:

$$(\nu_1, \ldots, \nu_H)^T \sim \text{Dirichlet}\Big(\frac{1}{H}, \ldots, \frac{1}{H}\Big). \tag{3.8}$$

Inoltre Rousseau & Mengersen (2011) mostrano che tale approssimazione favorisce l'eliminazione automatica delle classi ridondanti, in modo che la relativa distribuzione a posteriori si concentri sul vero numero di componenti H^0.

2. Il secondo livello di penalizzazione è specifico per ogni gruppo, e produce una rappresentazione a basso rango del vettore di probabilità associato. In particolare si cerca un distribuzione a priori Π_λ in grado di favorire l'eliminazione delle dimensioni non necessarie a caratterizzare $\pi^{(h)}$ secondo (3.4). Viene adattata al caso una distribuzione proposta da Bhattacharya et al. (2011) nello sviluppo di un modello gaussiano per fattori latenti; si tratta, fondamentalmente, di una produttoria di distribuzioni del tipo Gamma-Inversa, denotata come $\text{MIG}(a_1, a_2)$ ed espressa nel seguente modo

$$\lambda_r^{(h)} = \prod_{m=1}^{r} \frac{1}{\vartheta_m^{(h)}} \quad , \quad \vartheta_1^{(h)} \sim \text{Ga}(a_1, 1) \quad , \quad \vartheta_{m>1}^{(h)} \sim \text{Ga}(a_2, 1), \quad r = 1, \ldots, R \quad (3.9)$$

indipendentemente per ogni $h = 1, \ldots, H$. Gli elementi $\boldsymbol{\lambda}_r^{(h)}$ sono stocasticamente descrescenti verso 0 al crescere di r per valori di $a_2 > 0$, in modo da favorire una penalizzazione adattiva della rappresentazione sovra-parametrizzata di ognuno dei vettori di probabilità $\boldsymbol{\pi}^{(h)}(h = 1, \ldots, H)$, mentre il parametro a_1 invece controlla il grado totale di variabilità degli elementi di $\boldsymbol{\lambda}^{(h)}$. Si rimanda a Bhattacharya et al. (2011) per ulteriori proprietà teoriche.

Infine Π_Z e Π_X sono definite a priori come distribuzioni Gaussiane, rispettivamente:

$$\boldsymbol{Z} \sim N_{V(V-1)/2}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\mu} \in \Re^{V(V-1)/2}, \quad \boldsymbol{\Sigma} = diag(\sigma_1^2, \ldots, \sigma_{V(V-1)/2}^2)$$
$$\boldsymbol{X}_{vr}^{(h)} \sim N(0, 1), \quad v = 1, \ldots, V, \quad r = 1, \ldots, R, \quad h = 1, \ldots, H.$$

Per facilitare il calcolo della distribuzioni a posteriori, Durante et al. (2015) operano una riparametrizzazione del modello, ponendo $\bar{\boldsymbol{X}}^{(h)} = \boldsymbol{X}^{(h)}\boldsymbol{\Lambda}^{(h)1/2}$ e quindi $\boldsymbol{D^{(h)}} = \mathcal{L}(\bar{\boldsymbol{X}}^{(h)}\bar{\boldsymbol{X}}^{(h)T})$ per $h = 1, \ldots, H$. Data la distribuzioni a priori, si può campionare direttamente $\bar{\boldsymbol{X}}_{vr}^{(h)}|\lambda_{vr}^{(h)}$ da $N(0|\lambda_{vr}^{(h)})$ in maniera indipendente $v = 1, \ldots, V$, $r = 1, \ldots, R$ e $h = 1, \ldots, H$.

3.1.2 Distribuzione a posteriori

Il calcolo della distribuzione a posteriori è ottenuto adattando l'algoritmo di *data augmentation* proposto da Polson et al. (2013) per permettere di fare inferenza esatta in modelli bayesiani con verosimiglianza di tipo binomiale con funzione legame logistica. In generale, con il termine *data augmentation* si fa riferimento ad una tecnica computazionale per la costruzione di algoritmi di campionamento o di ottimizzazione iterative, basato su variabili latenti, allo scopo di introdurre una maggiore quantità di dati a quelli osservati, rendendoli più facili da analizzare. Il metodo fu reso popolare da Dempster et al. (1977) nell'algoritmo EM (*Expectation-maximization*) per la risoluzione di problemi di stima di massima verosimiglianza, mentre Tanner & Wong (1987) lo applicano in un contesto bayesiano nel calcolo della distribuzione a posteriori. L'idea alla base è piuttosto semplice ed è quella di trattare i dati a disposizione come realizzazioni di una variabile casuale latente Z della quale si conosce la distribuzione. Si immagina quindi una situazione in cui sia semplice simulare dalla distribuzione a posteriori dei parametri di interesse, condizionata non al campione osservato y, bensì a quello completo (y, z) di cui z sono realizzazioni da Z, e che sia agevole simulare la distribuzione predittiva $p(z|y, \theta)$ dei dati "mancanti".

La stima di un modello logistico dal punto di vista bayesiano è sempre risultata problematica, a causa della forma analitica della funzione di verosimiglianza associata. Polson

et al. (2013) presentano una soluzione a tale problematica analoga all'approccio di Albert & Chib (1993) sviluppato per il modello di tipo probit, che considera i dati binari come una censura dicotomica di una variabile aleatoria normale, e sfrutta un metodo di *data augmentation* per la stima del modello. Polson et al. (2013) sviluppano un approccio analogo basato sulla definizione di una nuova famiglia di distribuzioni Pólya-Gamma, della quale si riporta la definizione.

Definizione 3.1. Una variabile casuale X è detta avere una distribuzione Pólya-Gamma di parametri $b > 0$ e $c \in \mathbb{R}$, denotata come $X \sim PG(b, c)$ se

$$X \overset{\mathrm{D}}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k - 1/2)^2 + c^2/(4\pi^2)}, \tag{3.10}$$

dove $g_k \sim Ga(b, 1)$ sono variabili indipendenti con distribuzione Gamma, e con $\overset{\mathrm{D}}{=}$ indica l'uguaglianza in distribuzione.

Data questa nuova famiglia di distribuzioni, Polson et al. (2013) dimostrano che è possibile ottenere una rappresentazione della verosimiglianza binomiale in termini di *log-odds*, come mistura di distribuzioni di tipo normale. L'identità alla base di tale risultato è dato dalla seguente uguaglianza:

$$\frac{(e^{\psi})^a}{(1 + e^{\psi})^b} = 2^{-b} e^{\kappa\psi} \int_0^{\infty} e^{-\omega\psi^2/2} p(\omega) d\omega \tag{3.11}$$

dove $\kappa = a - b/2$ e $\omega \sim PG(b, 0)$. Posto $\psi = x^T\boldsymbol{\beta}$ come funzione di predittori linear, la funzione integranda è riconducibile al nucleo di una funzione di verosimiglianza normale rispetto $\boldsymbol{\beta}$, mentre la distribuzione condizionata di ω dato ψ rimane della famiglia Pólya-Gamma (si rimanda a Polson et al. (2013) per la dimostrazione).

La strategia di *data augmentation* è la medesima adottata da Albert & Chib (1993), con la differenza che le variabili latenti associate alle osservazioni, sono di tipo Pólya-Gamma. Sia y_i il numero di successi, n_i il numero di tentantivi e $x_i = (x_{i1}, \ldots, x_{ip})^T$ il vettore di regressori associati all'osservazione $i \in \{1, \ldots, N\}$. Sia quindi $y_i \sim Binom(n_1, 1/1 + e^{\psi_i})$, dove $\psi = x_i^T\boldsymbol{\beta}$ sono i *log-odds* delle probabilità di successo. Fissata per $\boldsymbol{\beta}$ una distribuzione a priori normale $\boldsymbol{\beta} \sim N(b, B)$, la distribuzione a posteriori è ottenuta iterando due passi:

$$(\omega_i|\boldsymbol{\beta}) \sim PG(n_i, x_i^T\boldsymbol{\beta})$$
$$(\boldsymbol{\beta}|y, w) \sim N(m_{\omega}, V_{\omega})$$

dove $V_{\omega} = (X^T\Omega X + B^{-1})^{-1}$ e $m_{\omega} = V_{\omega}(X^T\kappa + B^{-1}b)$, con $\kappa = (y_1 - n_1 x/2, \ldots, y_N - n_N/2)$ e $\Omega = diag(\omega_1, \ldots, \omega_N)$.

L'efficacia di tale strategia rispetto ad altri metodi proposti negli ultimi anni (Holmes & Held, 2006; Gramacy & Polson, 2012) sta nella possibilità di poter simulare in maniera efficente variabili casuali Pólya-Gamma, per mezzo di un algoritmo di accettazione-rifiuto.

Si rimanda invece a Choi & Hobert (2013) per la dimostrazione sull'ergodicità delle catena nella stima del modello.

3.1.3 Gibbs Sampling

L'algoritmo di *Gibbs Sampling* per il modello definito da Durante et al. (2015) consiste in due principali fasi in cui: (i) ogni osservazione $\mathcal{L}(\boldsymbol{A}_i), i = 1, \ldots, n$ è allocata in una delle H classi, secondo un processo di *stick-breaking* e quindi (ii) per ognuna delle classi occupate, si aggiornano le quantità $\boldsymbol{Z}, \boldsymbol{X}^{(h)}, \boldsymbol{\lambda}^{(h)}$, per $h = 1, \ldots, H$, attraverso una regressione logistica bayesiana per ognuna delle classi, come di seguito:

- Per ogni osservazione si estrae la classe latente di appartenenza con probabilità pari a

$$pr(G_i = h|-) = \frac{\nu_h \prod_{l=1}^{V(V-1)/2} \{\pi_l^{(h)}\}^{\mathcal{L}(A_i)_l} \{1 - \pi_l^{(h)}\}^{1-\mathcal{L}(A_i)_l}}{\sum_{m=1}^{H} \nu_m \prod_{l=1}^{V(V-1)/2} \{\pi_l^{(h)}\}^{\mathcal{L}(A_i)_l} \{1 - \pi_l^{(h)}\}^{1-\mathcal{L}(A_i)_l}} \tag{3.12}$$

  per ogni $h = 1, \ldots, H$ e $i = 1, \ldots, N$, e con $\boldsymbol{\pi}^{(h)}$ definita come in (3.4).

- Si aggiornano i pesi sulle classi

$$(v_1, \ldots, v_H)|- \sim \text{Dirichlet}\Big\{ \frac{1}{H} + \sum_{i=1}^{n} \mathbb{1}(G_i = 1), \ldots, \frac{1}{H} + \sum_{i=1}^{n} \mathbb{1}(G_i = H) \Big\}. \tag{3.13}$$

Secondo il modello definito da Durante et al. (2015), reti nella stessa classe sono tra loro indipendenti e identicamente distribuite, condizionatamente al vettore delle probabilità di connessione classe-specifico $\boldsymbol{\pi}^{(h)}, h = 1, \ldots, H$.
L'aggiornamento di tale vettore, e quindi delle quantità in cui è fattorizzato secondo (3.4), si ottiene adattando l'algoritmo di *data augmentation* tramite variabili Pólya-Gamma proposto da Polson et al. (2013), alle reti aggregate $\boldsymbol{Y}^1, \ldots, \boldsymbol{Y}^H$, in cui $\boldsymbol{Y}^{(h)} = \sum_{G_i = h} \mathcal{L}(\boldsymbol{A})$, per $h = 1, \ldots, H$.
Ridefinendo il modello per le reti aggregate, si ottiene:

$$(Y_l^{(h)}|\boldsymbol{Z}, \boldsymbol{X}^{(h)}, \boldsymbol{\lambda}^{(h)}) \sim \text{Binom}[n_h, 1/\{1 + \exp(-Z_l - \mathcal{L}(X^{(h)} \Lambda^{(h)} X^{(h)T})_l\}] \tag{3.14}$$

indipendentemente per $l = 1, \ldots, V(V-1)/2$ e $h = 1, \ldots, H$. Una volta allocate le reti nelle classi, l'algoritmo procede nel modo seguente:

- Si estraggono le variabili latenti Pólya-Gamma dalla distribuzione a posteriori

$$\omega_l^{(h)}|- \sim \text{PG}\Big\{ n_h, Z_l + \mathcal{L}(X^{(h)} \Lambda^{(h)} X^{(h)T})_l \Big\} \tag{3.15}$$

  per ogni classe $h = 1, \ldots, H$ e $l = 1, \ldots, V(V-1)/2$

- Si estrae il vettore di similarità Z dalla rispettiva distribuzione a posteriori

$$\boldsymbol{Z}|- \sim N_{V(V-1)/2}(\boldsymbol{\mu}_Z, \boldsymbol{\Sigma}_Z) \tag{3.16}$$

dove Σ_Z è una matrice $V(V-1)/2 \times V(V-1)/2$ con elementi diagonali $\sigma^2_{Z,l} = 1/(\sigma_l^{-2} + \sum_{h=1}^{H} \omega_l^{(h)})$ mentre μ_Z un vettore $V(V-1)/2$ di elementi $\mu_{Z,l} = \sigma^2_{Z,l}[\sigma_l^{-2}\mu_l + \sum_{h=1}^{H}\{Y_l^{(h)} - n_h/2 - \omega_l^{(h)}\mathcal{L}(X^{(h)}\Lambda^{(h)}X^{(h)T})\}]$

- La riparametrizzazione del modello come $\bar{\boldsymbol{X}}^{(h)} = \boldsymbol{X}^{(h)}\boldsymbol{\Lambda}^{(h)1/2}$ comporta che a priori $\bar{\boldsymbol{X}}^{(h)}$ sia distribuita come una $\mathrm{N}(0, \lambda_r^{(h)})$ condizionatamente a $\lambda_r^{(h)}$. È possibile quindi aggiornare ogni riga di $\bar{\boldsymbol{X}}^{(h)}$, $\bar{\boldsymbol{X}}_v^{(h)} = (\bar{X}_{v1}^{(h)}, \ldots, \bar{X}_{vR}^{(h)})^T$ condizionatamente a $\bar{\boldsymbol{X}}_{(-v)}^{(h)}$, dove $\bar{\boldsymbol{X}}_{(-v)}^{(h)}$ denota la matrice $(V-1)\times R$ ottenuta rimuovendo la v-esima riga di $\bar{\boldsymbol{X}}^{(h)}$, ridefinendo il problema in termini di regressione logistica bayesiana

$$\boldsymbol{Y}_{(v)}^{(h)} \sim \mathrm{Binom}(n_h, \boldsymbol{\pi}_{(v)}^{(h)}), \quad \mathrm{logit}(\boldsymbol{\pi}_{(v)}^{(h)}) = \boldsymbol{Z}_{(v)} + \bar{\boldsymbol{X}}_{(-v)}^{(h)T}\bar{\boldsymbol{X}}_{(v)}^{(h)}, \tag{3.17}$$

con $\boldsymbol{Y}_{(v)}^{(h)}$ e $\boldsymbol{Z}_{(v)}$ ottenute selezionando gli elementi di $Y_l^{(h)}$ e $Z_l^{(h)}$, rispettivamente, per ogni arco l corrispondente ai nodi $\{u, v\}$ tali che $u = v$ o $z = v$. con $u > z$, e riordinati secondo (3.17). Data la matrice $\boldsymbol{\Omega}_{(v)}^{(h)}$ i cui elementi diagonali sono i valori generati dalle corrispondenti variabili latenti con distribuzione Pólya-Gamma, la distribuzione a posteriori è ottenuta come:

$$* \bar{\boldsymbol{X}}_{(v)}^{(h)}|- \sim \mathrm{N}\Big\{\Big(\bar{\boldsymbol{X}}_{(-v)}^{(h)T}\boldsymbol{\Omega}_{(v)}^{(h)}\bar{\boldsymbol{X}}_{(-v)}^{(h)} + \boldsymbol{\Lambda}^{(h)-1}\Big)^{-1}\boldsymbol{\eta}_v^{(h)}, \Big(\bar{\boldsymbol{X}}_{(-v)}^{(h)T}\boldsymbol{\Omega}_{(v)}^{(h)}\bar{\boldsymbol{X}}_{(-v)}^{(h)} + \boldsymbol{\Lambda}^{(h)-1}\Big)^{-1}\Big\} \tag{3.18}$$

con $\boldsymbol{\eta}_v^{(h)} = \bar{\boldsymbol{X}}_{(-v)}^{(h)T}(\boldsymbol{Y}_{(v)}^{(h)} - \mathbf{1}_{V-1}n_h/2 - \boldsymbol{\Omega}_{(v)}^{(h)}\boldsymbol{Z}_{(v)})$

- Per ogni classe $h = 1, \ldots, H$, si estraggono gli elementi del vettore $\boldsymbol{\vartheta}^{(h)} = (\vartheta_1^{(h)}, \ldots, \vartheta_R^{(h)})$ caratterizzante la distribuzione $\mathrm{MIG}(a_1, a_2)$ per $\boldsymbol{\Lambda}^{(h)}$ dalla rispettiva distribuzione a posteriori:

$$\vartheta_1^{(h)}|- \sim \mathrm{Ga}\Big\{a_1 + \frac{VR}{2}, 1 + \frac{1}{2}\sum_{m=1}^{R}\theta_m^{-1}\sum_{v=1}^{V}(\bar{X}_{vm}^{(h)})^2\Big\}$$
$$\vartheta_{r>1}^{(h)}|- \sim \mathrm{Ga}\Big\{a_2 + \frac{V\times(R-r+1)}{2}, 1, +\frac{1}{2}\sum_{m=1}^{R}\theta_m^{-r}\sum_{v=1}^{V}(\bar{X}_{vm}^{(h)})^2\Big\},$$

dove $\theta_m^{-r} = \prod_{t=1,t\neq r}^{m}\vartheta_t^{(h)}$ per $r = 1, \ldots, R$, e si calcolano gli elementi $\lambda_r^{(h)}$ come definito in (3.9).

- Per ogni classe $h = 1, \ldots, H$ si calcola il corrispondente vettore delle probabilità di connessione $\boldsymbol{\pi}^{(h)}$ come $\boldsymbol{\pi}^{(h)} = \{1 + \exp[-\boldsymbol{Z}^{(h)} - \boldsymbol{D}^{(h)}]\}^{-1}$ con $\boldsymbol{D}^{(h)} = \mathcal{L}(\bar{\boldsymbol{X}}^{(h)}\bar{\boldsymbol{X}}^{(h)T})$

3.2 Modello a blocchi stocastici per la rete di primo livello

Il modello di Durante et al. (2015) fornisce una rappresentazione ridotta delle nostre reti di prodotti, in grado di tenere conto dell'intera struttura di dipendenze all'interno delle reti, e associa ogni rete ad una classe h. Per ogni classe è definito un diverso processo generativo di rete, che descrive la *forma* della rete; è di interesse verificare se la forma associata alle reti di secondo livello sia informativa sulle connessioni tra i nodi nel primo livello. A tale scopo si propone nel seguente paragrafo un modello a blocchi stocastici per la rete di primo livello, in cui viene inserita una dipendenza sui colori dei nodi, data dalla forma delle reti associate.

Riprendendo la notazione definita nel Capitolo 2 per il modello a blocchi stocastici, si denota con W la matrice di adiacenza della rete di primo livello $\mathcal{G}$ composta da N nodi, i quali appartengono a B differenti categorie, definite come *blocchi* o *colori*. Siano i blocchi rappresentati da una variabile casuale $\boldsymbol{y} = (y_1, \dots, y_N)$ associata ai vertici, con valori in $1, \dots, B$, tali che $y_i = k$ se il vertice i è di colore k per $i \in \{1, \dots, N\}$ e $k = 1, \dots, B$. Dato un vertice i di colore k ed uno j di colore l, la probabilità di connessione tra i due vertici può essere scritta come $p(w_{ij}|y_i = k, y_j = l) = \eta_{kl}$, in cui η_{kl} è la probabilità di connessione specifica delle classi, tra i due nodi. Data la colorazione di due nodi, e quindi la probabilità di connessione associata, si può definire la presenza di un arco tra i due come un'estrazione indipendente da una variabile casuale Bernoulliana con probabilità di successo dipendente dai colori dei blocchi. Tale ipotesi induce una rappresentazione ridotta di dimensioni $B \times B$ della matrice di adiacenza in termini probabilistici, definita come

$$\boldsymbol{\eta} = \begin{pmatrix} \eta_{11} & \eta_{12} & \cdots & \eta_{1B} \\ \eta_{21} & \eta_{22} & \cdots & \eta_{2B} \\ \vdots & \vdots & \ddots & \vdots \\ \eta_{B1} & \eta_{B2} & \cdots & \eta_{BB} \end{pmatrix} \tag{3.19}$$

Si assume che i colori y_i siano parametri non noti, la cui distribuzione è dipendente dall'indicatore di classe G_i secondo $p(y_i = k|G_i = h) = \psi_{kh}$, ed è possibile definire una rappresentazione della probabilità condizionata di appartenenza ad un blocco, tramite la seguente matrice di *forma-colore*

$$\Psi_{b,h} = \begin{pmatrix} \psi_{11} & \psi_{12} & \cdots & \psi_{1B} \\ \psi_{21} & \psi_{22} & \cdots & \psi_{2B} \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{H1} & \psi_{H2} & \cdots & \psi_{HB} \end{pmatrix} \tag{3.20}$$

Data la forma h della i-esima rete di secondo livello, e quindi la rispettiva riga della matrice Ψ, la colorazione del nodo i è una misura di probabilità discreta definita su B blocchi data da

$$p(y_i|G_i = h) = \sum_{k=1}^{B} \psi_{kh} \delta_{\{y_i = k\}}. \tag{3.21}$$

Come per il modello delle reti di secondo livello, si definisce come distribuzione a priori sui blocchi un processo di Dirichlet, condizionatamente alla forma della rete associata al nodo, in modo da inferire dai dati il numero e la composizione dei blocchi. Approssimando il processo di Dirichlet come in (3.8), si definiscono H distribuzioni a priori sui pesi associati ai blocchi:

$$(\psi_{h1}, \dots, \psi_{hB}|\boldsymbol{G}) \sim Dirichlet\Big(\frac{1}{B}, \dots, \frac{1}{B}\Big), \quad h = 1, \dots, H \tag{3.22}$$

Dati sia la forma che il colore dei nodi, la funzione di verosimiglianza di $\mathbf{x}$ condizionata a $\mathbf{y}$ e $\boldsymbol{\eta}$, è data dal prodotto di variabili casuali Binomiali indipendenti:

$$p(\mathbf{w}|\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\Psi}) = \prod_{0 \le k \le l \le B} \eta_{kl}^{f_{kl}} (1 - \eta_{kl})^{n_{kl} - f_{kl}} \tag{3.23}$$

dove

$$f_{kl} = \frac{1}{1 + \delta_{kl}} \sum_{1 \le i \ne j \le N} w_{ij} \mathbb{1}(w_i = k) \mathbb{1}(w_j = l) \tag{3.24}$$

denota il numero di archi presenti nel grafo che uniscono un vertice di colore k ad un'altro di colore l, mentre $\delta_{kl} = 1$ per $k = l$ e $\delta_{kl} = 0$ per $k \ne l$. L'elemento n_{kl} denota il numero totale di vertici all'interno dei gruppi:

$$n_{kl} = \begin{cases} n_k n_l & \text{if } k \ne l \\ \binom{n_k}{2} & k = l \end{cases} \tag{3.25}$$

Marginalizzando le probabilità sui blocchi, si ottiene infine la seguente distribuzione congiunta per $(\mathbf{x}, \mathbf{y})$

$$p(\mathbf{w}, \mathbf{y}|\boldsymbol{\eta}, \boldsymbol{\Psi}) = \prod_{b=1}^{B} \prod_{h=1}^{H} \psi_{bh} \prod_{0 \le k \le l \le B} \eta_{kl}^{f_{kl}} (1 - \eta_{kl})^{n_{kl} - f_{kl}} \tag{3.26}$$

3.2.1 Stima del modello

Una volta determinata la forma delle reti di secondo livello , secondo il modello definito in Durante et al. (2015), la distribuzione a posteriori del ottiene in maniera agevole, aggiungendo i seguenti passi al *Gibbs Sampling* della sezione 3.1.3

- Ogni nodo è associato ad un blocco, tramite estrazione dalla distribuzione a posteriori sui blocchi, le cui probabilità sono calcolate come

$$p(y_i = k|\{y_j\}_{j \ne i}, \boldsymbol{\eta}, \boldsymbol{w}, G_i) = \frac{\psi_{kh} \prod_{l=0}^{B} (\eta_{kl})^{d_{il}} (1 - \eta_{kl})^{n_l - d_{il}}}{\sum_{k=0}^{B} \psi_{kh} \prod_{l=0}^{B} (\eta_{kl})^{d_{il}} (1 - \eta_{kl})^{n_l - d_{il}}} \tag{3.27}$$

  dove $d_{il} = \sum_{1 \le j \le N, j \ne i} x_{ij} \mathbb{1}(y_j = l)$ for $l = 0, \dots, B$ è il numero di vertici appartenenti alla classe l che sono connessi all'i-esimo nodo, mentre n_l il numero totale di nodi nel l-esimo blocco.

- Fissata una distribuzione a priori $Beta(1,1)$ per ogni probabilità di connessione η_{kl}, i conteggi f_{kl} definiti in (3.24) sono un'estrazione dalla rispettiva distribuzione coniugata binomiale di probabilità η_{kl}; si ha quindi la seguente distribuzione a posteriori per η_{kl}

$$\begin{aligned} &f_{kl} \sim Bin(n_{kl}, \eta_{kl}) \quad \eta_{kl} \sim Beta(1,1) \\ &(\eta_{kl} | \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{\Psi}) \sim Beta(1 + f_{kl}, 1 + n_{kl} - f_{kl}) \end{aligned}$$

- Si aggiornano i pesi sui blocchi, condizionatamente alla classe latente h

$$(\psi_{h1}, \ldots, \psi_{hB} | \boldsymbol{G}, -) \sim Dirichlet\Big(\frac{1}{B} + c_{1h}, \ldots, \frac{1}{B} + c_{Bh}\Big) \tag{3.28}$$

  dove $c_{bh} = \sum_{i=1}^{N} \mathbb{I}\{y_i = b\}\mathbb{I}\{G_i = h\}$

Capitolo 4

Studio di simulazione

4.1 Simulazione di una rete di reti

Si è condotto uno studio di simulazione con l'obiettivo di valutare le prestazioni del modello proposto. I dati sono stati generati in modo da imitare il processo probabilistico assunto dal modello; fissate le H^0 classi per le reti di secondo livello, sono state generate le osservazioni di rete secondo il modello di Durante et al. (2015); ad ogni classe di rete h, si è associato un colore di nodo k nella rete di primo livello. Definite le probabilità di connessione intra-blocco e tra i blocchi, si sono generate le connessioni della rete di primo livello. In particolare si sono considerate $H^0 = 5$ classi (forme) per le reti di secondi livello, associando ogni classe uno dei $B^0 = 5$ blocchi (colori) nella rete di primo livello. Si è considerata una numerosità campionaria pari a 100, ed un numero di nodi pari a 20 per le reti di secondo livello.

Le reti di secondo livello sono associate alla rappresentazione del vettore di probabilità di connessione definito in (3.4): i *log-odds* dei vettori di probabilità $\boldsymbol{\pi}^{0(h)}$ sono scomposti in una componente di similarità $\boldsymbol{Z}$ e una componente di deviazione classe-specifica $\boldsymbol{D}^{(h)}$. Il livello di complessità all'interno di ogni classe è regolato da quest'ultima componente, ulteriormente fattorizzata come $\mathcal{L}(\boldsymbol{X}^{(h)}\boldsymbol{\Lambda}^{(h)}\boldsymbol{X}^{(h)T})$, dove $\boldsymbol{\Lambda}$ è una matrice $R \times R$ di elementi $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_R)$. Il vettore $\boldsymbol{\lambda}$ regola la complessità della struttura di rete spcifica di classe, in quanto la sua dimensione definisce quella dello spazio latente $R^{(h)0}$ associato ai vettori di probabilità di connessione di ogni classe. Nella simulazione si è fissato

$$\boldsymbol{\lambda}^{0(h)} = \begin{cases} (5,0,0)^T & (R^{0(h)}=1) \quad h = 1,\ldots,20, \quad k=1 \\ (6,2,0)^T & (R^{0(h)}=2) \quad h = 21,\ldots,40, \quad k=2 \\ (3.5,1,0)^T & (R^{0(h)}=2) \quad h = 41,\ldots,60, \quad k=3 \\ (2.5,1.5,1)^T & (R^{0(h)}=3) \quad h = 61,\ldots,80, \quad k=4 \\ (10,7,2)^T & (R^{0(h)}=3) \quad h = 81,\ldots,100, \quad k=5 \end{cases} \tag{4.1}$$

in cui $R^{0(h)}$ indica lo spazio latente specifico di classe del vettore di probabilità di connessione. Gli elementi delle matrici $\boldsymbol{X}^h$ sono generate come normali standard, mentre si è considerato nullo il vettore di similarità $\boldsymbol{Z}$.

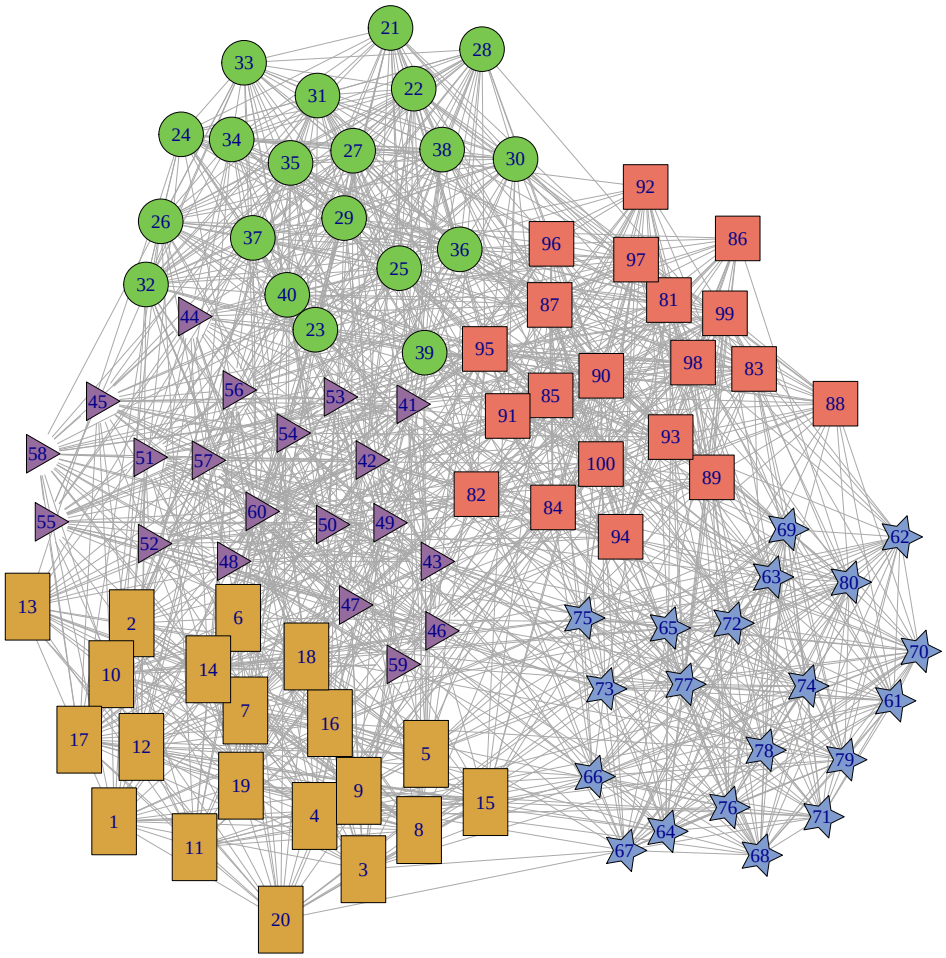

Figura 4.1: Rete generata nella simulazione: ad ogni forma di rete di secondo livello corrisponde un colore nella rete di primo livello

Ad ognuna delle classi di rete di secondo livello, è stato associato un colore (blocco) al nodo corrispondente nella rete di primo livello; si sono quindi fissate le probabilità di connessione intra-blocco e tra i blocchi, in modo che la probabilità di connessione tra i vertici appartenenti allo stesso blocco fosse maggiore. Si è quindi simulata la rete di primo livello estraendo, per ogni coppia di nodi, il relativo arco da una variabile casuale bernoulliana con probabilità di successo data dal colore dei nodi. In figura Figura 4.1 si riporta il grafico della rete simulata.

La distribuzione a posteriori è stata calcolata secondo il modello esposto nel Capitolo 3, utilizzando dei limiti superiori di tipo conservativo $H = 30$, $R = 10$, per le reti di secondo livello e $B = 20$ per la rete di primo livello. Le distribuzioni a priori per i parametri della rete di primo livello sono state scelte di tipo non informativo come descritto nella Paragrafo 3.2.1, mentre per i parametri delle reti di secondo livello si è tenuto conto dei suggerimenti

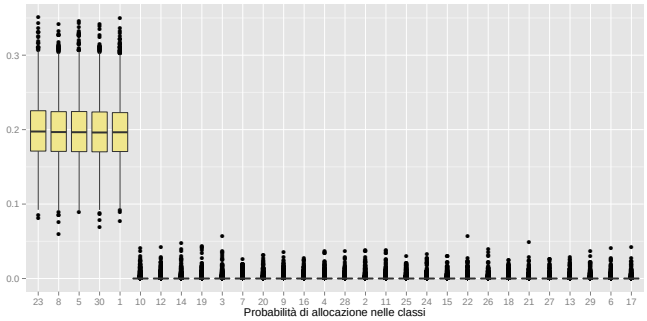

Figura 4.2: Distribuzione della probabilità di allocazione nelle classi latenti per le reti di secondo livello, riordinate in ordine decrescende secondo le medie a posteriori di ν_h

in Durante et al. (2015): allo scopo di favorire l'eliminazione delle dimensioni ridondanti all'interno di ogni classe, si sono fissati i parametri $a_1 = 2.5$ e $a_2 = 1.5$ per la distribuzione a priori $MIG(a_1, a_2)$, mentre si è scelta una distribuzione a priori sparsa per le componenti del vettore di similarità $\boldsymbol{Z}$, fissando $\mu_0 = 0$ e $\sigma_0^2 = 10$.

4.2 Stima e *label-switching*

Sono state effettuate 5000 iterazioni dall'algoritmo di *Gibbs Sampling* ed è stato fissato un *burn-in* di 1000. Un problema tipico nella stima del modelli di mistura di tipo bayesiano non parametrico con algoritmi di tipo MCMC, è quello definito col termine di *label-switching* (Redner & Walker, 1984) e descrive situazioni in cui la distribuzione della funzione di verosimiglianza del modello risulta invariante rispetto a permutazioni di etichetta delle componenti di mistura. Nell'analisi bayesiana, tale invarianza produce spesso delle distribuzioni a posteriori simmetriche o multimodali per i parametri dipendenti dalla componente: una situazione simile risulta problematica quando l'inferenza è classe specifica, dato che le medie a posteriori dei parametri relativi, risultano simmetriche e quindi non informative. In altre parole, ad ogni iterazione dell'algoritmo di stima, le unità vengono allocate insieme alla stessa classe seppure l'etichetta non è la stessa: è possibile ottenere la giusta composizione delle classi stimandola come massimo a posteriori della distribuzione, nel caso non sia di interessa l'inferenza sui parametri specifici di classe. Se invece si è interessate a condurre un'analisi di inferenze su tali parametri, una possibile soluzione al problema è data da algoritmi di *relabeling*, come quello proposto da Stephens (2000).

Il modello proposto non presenta una situazione di *label-switching* per quanto riguarda le reti di secondo livello; in Figura 4.2 è anche possibile visualizzare l'effetto di penalizzazione della distribuzione a priori Π_ν della probabilità di allocazione alle classi. Le classi non necessarie a descrivere i dati, vengono via via lasciate vuote, e la rispettiva probabilità di allocazione cala fino ad essere prossima allo zero.

La problematica del *label-switching* sorge nella definizione dei blocchi nella rete di primo livello; tuttavia non siamo interessati a fare inferenza sulle probabilità di connessione specifica di blocco ma piuttosto sulla composizione dei gruppi e la loro caratterizzazione in termini di classi di reti di secondo livello. Stimando le classi e blocchi come massimo a posteriori delle

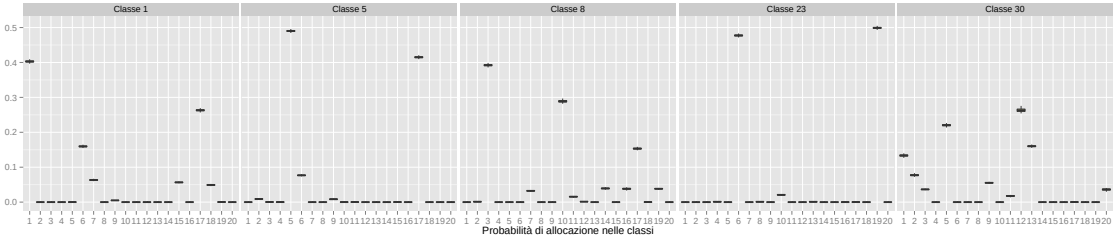

Figura 4.3: Distribuzione della probabilità di allocazione ai blocchi per ogni classe: si può notare l'effetto del fenomeno di *label-switching*: vi sono dei blocchi con probabilità di allocazione molto vicine. Dato che le unità si "muovono assieme" nelle classi, è possibile ottenere la stima esatta come massimo a posteriori.

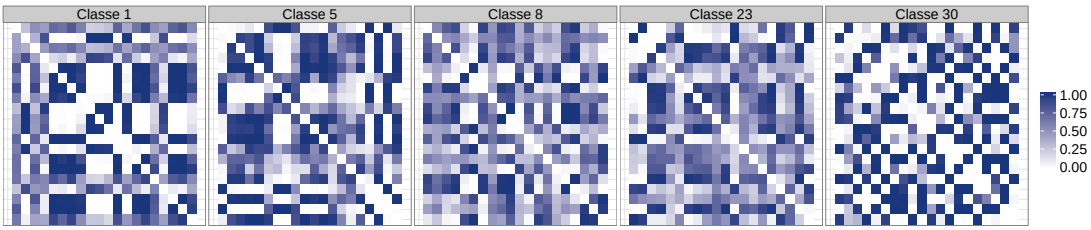

Figura 4.4: Per le 5 classi simulate: nella parte triangolare inferiore si ha la media a posteriori delle probabilità di connessione, mentre la parte triangolare superiore mostra le vere probabilità generate.

relative distribuzioni, otteniamo la giusta allocazione definita in partenza, come mostrato in Tabella 5.1. Possiamo notare l'effetto del *label-switching* nella Figura 4.3; siccome la composizione dei blocchi rimane invariante rispetto a permutazioni delle etichette, è possibile ottenere le giuste classificazioni. Si può anche notare lo stesso effetto di penalizzazione osservato per le classi di secondo livello, che elimina i blocchi ridondanti.

	Blocco 1	Blocco 12	Blocco 19	Blocco 3	Blocco 5
Classe 1	20	0	0	0	0
Classe 5	0	0	0	0	20
Classe 8	0	0	0	20	0
Classe 23	0	0	20	0	0
Classe 30	0	20	0	0	0

Tabella 4.1: Numerosità stimate per blocchi e classi; il modello riesce ad associare correttamente ogni classe ad un unico blocco

In Figura 4.2 si mostrano invece le medie a posteriori delle probabilità di connessione specifiche di classe $\pi^{h(0)}$ ottenute dal modello triangolare inferiore), in confronto con le vere probabilità generate (matrice triangolare superiore). Si può notare che le probabilità di connessione vengono stimate in maniera corretta.

Capitolo 5

Applicazione ai dati

I dati disponibili riguardano 135 agenzie assicurative e 15 tipologie di prodotto, riespressi all'interno di una struttura a due livelli di reti: un primo livello rappresentante le agenzie in termini di similarità (agenzie simili sono connesse), e un secondo livello in cui per ogni agenzia è specificata una rete di prodotti, descrivente la politica di vendita dell'agenzia. L'obiettivo che ci siamo posti è quello di ottenere una rappresentazione ridotta di tale struttura secondo dei raggruppamenti di agenzie, caratterizzate a loro volta da insiemi di politiche di vendita. A tale scopo è stato presentato nel Capitolo 3 un modello bayesiano non parametrico per una rete di reti, e si riportano nel presente capitolo i risultati ottenuti.

5.1 Discussione sulla stima del modello

Il modello presentato nel Capitolo 3 prevede una stima per mezzo di un algoritmo di *Gibbs Sampling*: la convergenza dell'algoritmo non dipende dalle assegnazioni iniziali delle unità in classi e in blocchi, ma tale scelta impatta invece sulla velocità della convergenza; nella pratica può risultare quindi utile inizializzare alcuni parametri in maniera efficiente, soprattutto se le classi e i blocchi non risultano ben separati. In particolare si sono inizializzate le classi delle reti di secondo livello secondo una procedura di *clustering* gerarchico basata sulla distanza di Jaccard con metodo di Ward (Murtagh, 1985), dividendo il campione in H classi. Sempre per le reti di secondo livello, si è fissato il vettore delle similarità $\boldsymbol{Z}$ pari al *log-odds* delle frequenze degli archi osservate nel campione di reti. Per inizializzare i blocchi della rete di primo livello, si è applicato il metodo di Louvain Blondel et al. (2008) per l'individuazione delle comunità nei grafi, e inizializzato i blocchi con i risultati ottenuti.

Si è calcolato il modello attraverso 10000 iterazioni dell'algoritmo di *Gibbs Sampling*, fissando gli stessi valori utilizzati nella simulazione del Capitolo 4 per le distribuzioni a priori, e stimato le quantità di interesse tenendo conto di un *burn-in* di 3000 iterazioni. Si sono ispezionati tutti i *trace-plots* delle probabilità di connessione tra i prodotti per valutare la convergenza del modello, e si riporta a titolo esemplificativo in Figura 5.1 il *trace-plot* della probabilità di connessione tra il prodotto Casa e il prodotto Investimento all'interno di ognuna delle classi di rete definite. Le catene raggiungono la convergenze e si ha un discreto *mixing* considerato l'elevato numero di parametri coinvolti nel processo di stima. Le classi

e i blocchi sono state stimate come massimo a posteriori della relativa usando come stima il massimo a posteriori (MAP) delle distribuzioni sulle classi.

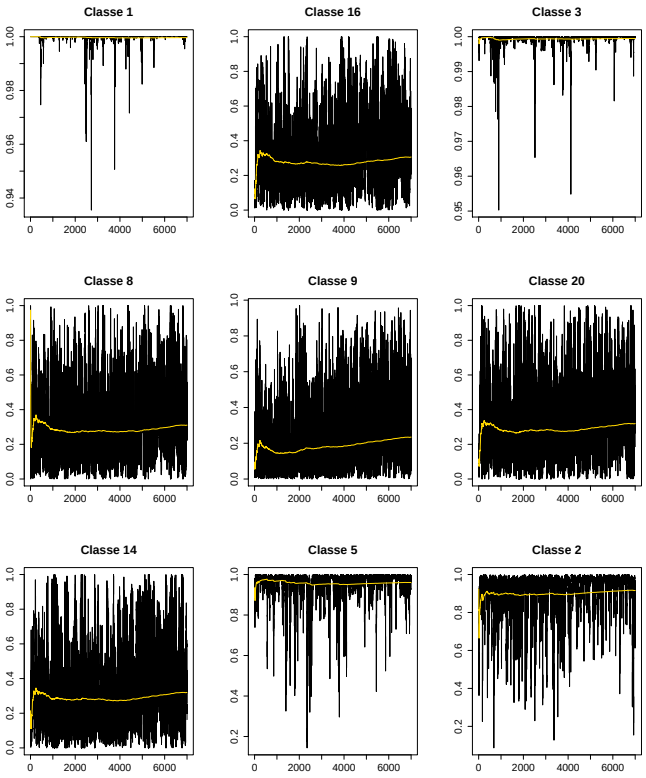

Figura 5.1: Per le classi non vuote, i *trace-plots* della probabilità $\pi^h_{3,10}$ di connessione tra il nodo prodotto Casa e il nodo prodotto Investimento; in giallo la stima cumulata delle relative medie.

5.2 Risultati

In Tabella 5.1 si riportano le numerosità per blocco e rete, mentre in Figura 5.2 si rappresenta la media a posteriori delle probabilità di allocazione ai blocchi date le classi delle reti di secondo livello, per le classi/blocchi occupati. Si sono ottenuti tre principali blocchi, e un blocco con un una singola unità (Blocco 15 in Figura 5.7).

	Blocco 1	Blocco 15	Blocco 2	Blocco 3
Classe 1	0	0	0	27
Classe 14	14	0	0	0
Classe 16	13	0	0	0
Classe 2	0	0	0	1
Classe 20	4	0	0	0
Classe 3	27	0	0	0
Classe 5	0	1	0	0
Classe 8	0	0	23	0
Classe 9	0	0	25	0

Tabella 5.1: Numerosità stimate per blocco e classe

In Figura 5.3 si ha una rappresentazione della rete delle agenzie colorando i nodi a seconda dei blocchi, e fissando una forma per i vertici a secondo della classe nel blocco; la posizione

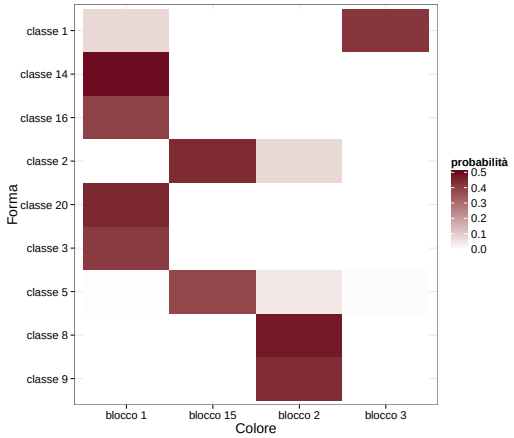

Figura 5.2: Distribuzione delle probabilità di allocazione al blocco data la classe di appartenenza, per le classi e i blocchi non vuoti

dei nodi è la stessa della rappresentazione nel Capitolo 1. Il modello coglie delle strutture di similarità tra agenzie diverse da quelle dell'algoritmo di rappresentazione, in quanto subisce l'effetto delle covariate.

Si è ottenuta quindi una rappresentazione a blocchi della rete di agenzie e, per ogni blocco, una caratterizzazione in classi delle relative reti di prodotti come tipologie di politica di vendita. Le strutture di dipendenza comuni alle reti di secondo livello, sono descritte dal valore atteso $\bar{\boldsymbol{\pi}} = \mathrm{E}\{\mathcal{L}(\boldsymbol{\mathcal{A}})\} = \sum_{\alpha \in \mathcal{Y}_V} \boldsymbol{\alpha} p_{\mathcal{L}(\boldsymbol{\mathcal{A}})}(\alpha)$ calcolabile come $\bar{\boldsymbol{\pi}} = \sum_{h=1}^{H} \nu_h \boldsymbol{\pi}^{(h)}$. Si rimanda a Durante et al. (2015) per la dimostrazione. È di maggiore interesse invece valutare le deviazioni classe-specifiche delle probabilità di connessione calcolando le medie a posteriori delle differenze $\boldsymbol{\pi}^{(h)} - \bar{\boldsymbol{\pi}}$, per $h = 1, \ldots, H$. Nelle figure seguenti sono rappresentate tali quantità per le classi non vuote, come distribuzioni di probabilità sulla matrice di adiacenza raggruppate secondo il blocco di appartenenza. Per ogni coppia di prodotti, il corrispondente quadrato nella matrice descrive quindi quanto è più o meno probabile che i due prodotti siano connessi per quella classe; in altre parole si ha una descrizione in termini probabilistici delle tipologia di politica di vendita in termini di clienti pluri-prodotto delle agenzie simili tra loro.

Il blocco 1 (Figura 5.4) è quello risultato più numeroso e con più classi di reti di prodotto: le classi 3 e 16, sono relative ad agenzie con una composizione di clienti pluri-prodotto più simile a quello complessivo, mentre vi sono sono forti deviazioni per quanto riguarda le agenzie delle classi 14 e 20. Analogamente nei blocchi 2 (Figura 5.5) e 3 (Figura 5.6)si ha una divisione tra agenzie in linea con la politica generale e quelle con forti deviazioni. Differente è il caso del blocco 15 (Figura 5.7), che contiene un'unica agenzia con una rete di prodotto non assimilabile alle altre, che presenta una probabilità di connessione del prodotto di tipo Casa molto minore rispetto a quella complessiva.

Data la composizione delle classi dei blocchi, ovvero dei gruppi di agenzie e la caratterizzazione di tali gruppi in termini di politiche di vendita, è possibile determinare due strategie principali per una campagna di marketing per i clienti pluri-prodotto:

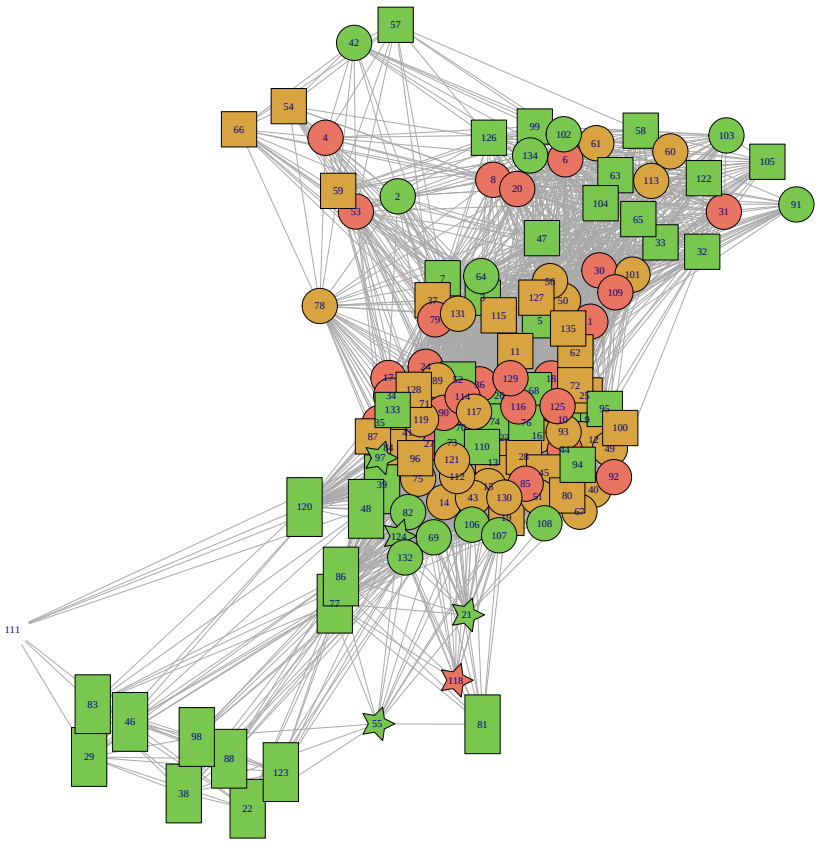

Figura 5.3: Rappresentazione delle forme e dei blocchi stimati, sulla rete iniziale; si noti che le forme di vertici si ripetono tra in nodi di colore diverso, ma non rappresentano la stessa classe; si è decido di utilizzare gli stessi simboli grafici allo scopo di non appesantire il grafico.

- incrementare delle vendite di coppie di prodotti con maggiore probabilità di connessione, che necessita un impiego di risorse minore per quelle classi che presentano già una buona base di connessioni di prodotto (classi 3, 16, 8), differenziando la composizione delle coppie in base al blocco di appartenenza.

- allineare all'andamento generale le agenzie che invece presentano forti deviazioni in termini di connessioni, in base alla considerazione che agenzie nello stesso blocco hanno lo stesso potenziale, in quanto simili tra loro: ad esempio, si può pensare che le classi nel blocco 1 rappresentino una diversa fase di passaggio di evoluzione delle politiche di vendita della agenzie. Dalla classe 14 alla 20 e alla 16 le deviazioni in termini di probabilità di connessione sono via via meno accentuate; idealmente si vorrebbe che le agenzie si spostino man mano nella classe 3.

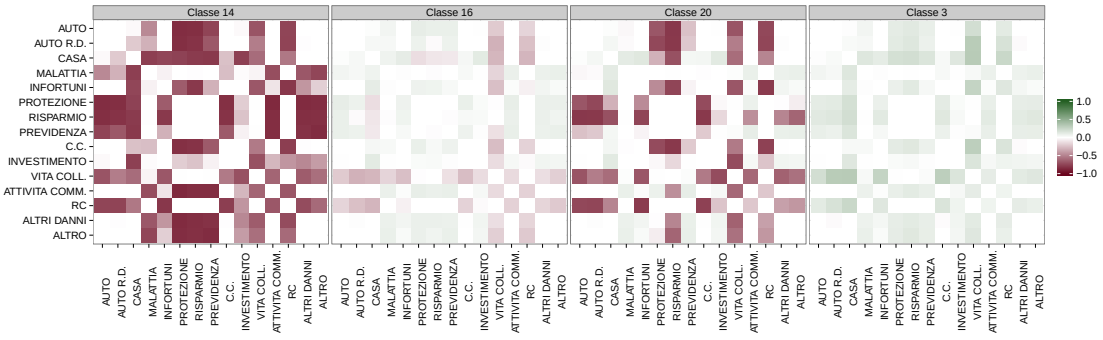

Figura 5.4: Blocco 1: matrice di adiacenza delle deviazioni delle probabilità di connessione $\boldsymbol{\pi}^{(h)} - \bar{\boldsymbol{\pi}}$, per h =14 (14 agenzie), 16 (13 agenzie), 20 (4 agenzie), 3 (27 agenzie)

La rappresentazione ottenuta dal modello, fornisce anche una possibile strategia di valutazione degli effetti di *trade-marketing* nel tempo; data la composizione dei gruppi e delle classi al mese t, la si può confrontare con quella del mese precedente. Ad esempio, si può monitorare la distribuzione delle agenzie nelle classi e nei blocchi, e verificare se vi è un'"evoluzione" delle politiche di vendita delle agenzie; oppure è possibile anche effettuare un confronto solo in termini generali considerando le strutture di dipendenza comuni descritte dal vettore $\bar{\boldsymbol{\pi}}$. In conclusione, il modello stimato costituisce una buona base di partenza per la definizione di azioni di *trade marketing* specifica per gruppi di agenzie, in modo da fornire un compromesso tra una campagna specifica per agenzia ed una del tutto generica. Inoltre la rappresentazione ottenuta della distribuzione delle probabilità di connessione, presenta il vantaggio di essere sia di facile comprensione che accattivante da un punto di vista grafico, e quindi con un buon potenziale di inserimento in un contesto aziendale.

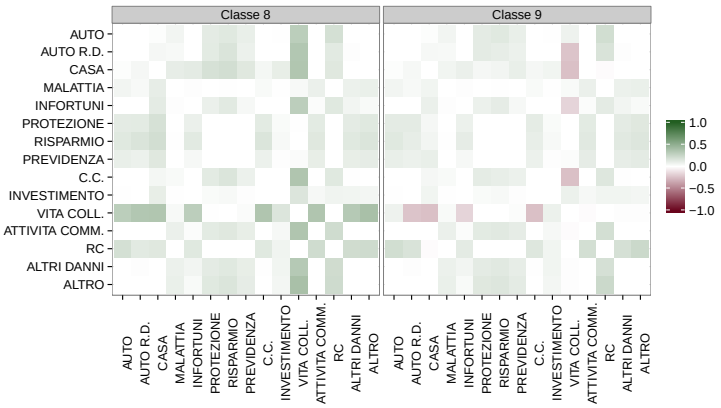

Figura 5.5: Blocco 2: matrice di adiacenza delle deviazioni delle probabilità di connessione $\boldsymbol{\pi}^{(h)} - \bar{\boldsymbol{\pi}}$, per h = 8 (23 agenzie),9 (25 agenzie)

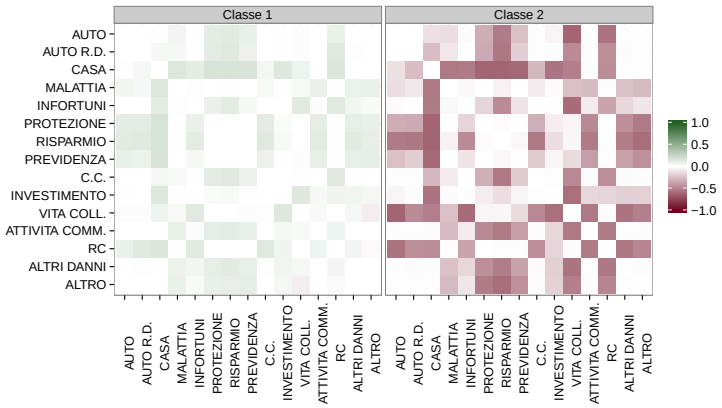

Figura 5.6: Blocco 3: matrice di adiacenza delle deviazioni delle probabilità di connessione $\boldsymbol{\pi}^{(h)} - \bar{\boldsymbol{\pi}}$, per $h = 1$(27 agenzie),2 (1 agenzia)

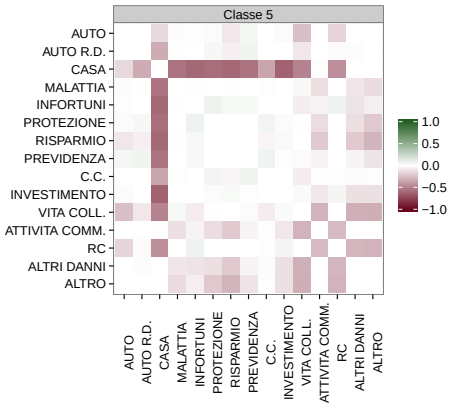

Figura 5.7: Blocco 15: matrice di adiacenza delle deviazioni delle probabilità di connessione $\boldsymbol{\pi}^{(h)} - \bar{\boldsymbol{\pi}}$, per $h = 5$ (1 agenzia)

Conclusioni

Nella presente tesi si è proposta una metodologia di analisi alternativa del paradigma di vendita di una compagnia assicurativa, finalizzata alla definizione di campagne di *trade marketing*. Si è messo in luce come tale paradigma possa essere rappresentato in termini di una struttura di rete a due livelli e proposto un modello bayesiano non parametrico in grado di descrivere appieno entrambe le strutture, attraverso un procedimento di *clustering* in due livelli.

Una volta colto il meccanismo di rappresentazione dei dati come una rete di reti, il modello proposto ha la capacità di rappresentare una struttura complessa in maniera accattivante e di facile interpretazione; blocchi di agenzie sono caratterizzate da alcune forme di reti di prodotti rappresentabili graficamente come probabilità sulle matrici di adiacenza. Inoltre fornisce una possibile strategia in grado di monitorare nel tempo gli effetti delle campagne avviate.

Al di fuori del contesto aziendale, strutture di rete di reti sono presenti in altri ambiti di ricerca, quali le neuroscienze, la biologia o il reperimento dell'informazione. Come per l'analisi di rete "semplice" si sono definiti problemi diversi e adattamenti al caso in base all'ambito di applicazione, allo stesso modo vi è spazio per il modello proposto.

L'analisi statistica di una struttura di rete di reti è una problematica di ricerca per la quale non è stato ancora definite delle metodologie rodate; come per tutti i modelli embrionali, il nostra modello presenta dei margini di miglioramento ed estensione. Una possibilità è quella di rendere stabile il processo di stima, in modo da permettere un inferenza sui parametri specifici per i blocchi: a tale scopo è necessario gestire il fenomeno di *label switching* per mezzo di algoritmi di *relabeling*. Un'estensione del modello invece, consiste nell'affiancare all'indice di classe, altri attributi associati ai nodi della rete di primo livello, possibile attraverso l'uso di una regressione bayesiana multinomiale, come quella definita in Polson et al. (2013).

Bibliografia

Albert, J. H. & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association* **88**, 669–679.

Azzalini, A. & Scarpa, B. (2012). *Data Analysis and Data Mining: An Introduction*. Oxford University Press.

Bhattacharya, A., Dunson, D. B. et al. (2011). Sparse bayesian infinite factor models. *Biometrika* **98**, 291.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**, P10008.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology* **2**, 113–120.

Choi, H. M. & Hobert, J. P. (2013). The polya-gamma gibbs sampler for bayesian logistic regression is uniformly ergodic. *Electron. J. Statist.* **7**, 2054–2064.

Dai, B., Ding, S., Wahba, G. et al. (2013). Multivariate bernoulli distribution. *Bernoulli* **19**, 1465–1483.

Daudin, J.-J., Picard, F. & Robin, S. (2008). A mixture model for random graphs. *Statistics and computing* **18**, 173–183.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)* , 1–38.

Durante, D., Dunson, D. B. & Vogelstein, J. T. (2015). Nonparametric Bayes Modeling of Populations of Networks. *ArXiv e-prints* .

Erdős, P. & Rényi, A. (1959). On random graphs. I. *Publ. Math. Debrecen* **6**, 290–297.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–230.

Fienberg, S. E. & Wasserman, S. S. (1981). Categorical data analysis of single sociometric relations. *Sociological methodology* , 156–192.

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

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry* , 35–41.

Fruchterman, T. M. & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Softw., Pract. Exper.* **21**, 1129–1164.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* **27**, 857–871.

Gramacy, R. B. & Polson, N. G. (2012). Simulation-based regularized logistic regression. *Bayesian Anal.* **7**, 567–590.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition.* Springer Series in Statistics. Springer.

Holland, P. W., Laskey, K. B. & Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social networks* **5**, 109–137.

Holmes, C. C. & Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Anal.* **1**, 145–168.

Ishwaran, H. & Zarepour, M. (2002). Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica* **12**, 941–963.

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

Murtagh, F. (1985). Multidimensional clustering algorithms. *Compstat Lectures, Vienna: Physika Verlag, 1985* **1**.

Newman, M. E. (2003). Mixing patterns in networks. *Physical Review E* **67**, 026126.

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

Polson, N. G., Scott, J. G. & Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association* **108**, 1339–1349.

Redner, R. A. & Walker, H. F. (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM review* **26**, 195–239.

Rousseau, J. & Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**, 689–710.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika* **31**, 581–603.

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica Sinica* **4**, 639–650.

Snijders, T. A., Pattison, P. E., Robins, G. L. & Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociological methodology* **36**, 99–153.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**, 795–809.

Tallberg, C. (2004). A bayesian approach to modeling stochastic blockstructures with covariates. *Journal of Mathematical Sociology* **29**, 1–23.

Tanner, M. A. & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association* **82**, 528–540.