Università di Pisa

Facoltà di Economia
Facoltà di Scienze
Matematiche, Fisiche e Naturali

Dipartimento di Informatica
Corso di Laurea Magistrale in Informatica
per l'Economia e per l'Azienda
(Business Informatics)

Tesi di Laurea

Scoperta di segregazione su reti sociali economiche

Candidato:

Alessandro Baroni

Relatore:

Prof. Salvatore Ruggieri

Anno Accademico 2011/2012

Sommario

La storia dell'umanità è costellata di eventi che indicano chiari trattamenti non paritari attuati nei confronti di individui o gruppi di individui, definiti in letteratura *gruppi protetti*. Questo fenomeno si chiama discriminazione ed è strettamente collegato ed a tratti commistionato con la segregazione.

Nel mercato del lavoro, sensibile a questi fenomeni vietati dal *D.Lgs. 198/2006*, urge la necessità di capire chiaramente in che intensità si manifestino e per quali gruppi protetti.

L'obiettivo della tesi è l'ideazione, la progettazione e lo sviluppo di un sistema per la misura della segregazione all'interno delle comunità che compongono una rete sociale economica.

Una rete sociale economica è una rete di aziende collegate per mezzo di amministratori a loro comuni.

Nella tesi si descrive lo studio della letteratura finalizzato all'individuazione di una metodologia scientifica specifica e le tecniche implementate per estrarre dai dati la conoscenza necessaria.

É stato utilizzato un approccio sperimentale differente dalle tecniche classiche che coinsiste nel mostrare i risultati in funzione di molteplici gruppi protetti per dare profondità al quadro prospettico.

Il sistema è stato sviluppato per misurare l'indice di isolamento e l'entropia, per più gruppi protetti contemporaneamente. Tale metodologia permette, a parità di analisi, migliori performance.

Se in letteratura si definiscono, attraverso stretti vincoli, pochi gruppi protetti da cui estrarre informazione, nel nostro lavoro sono stati rilassati alcuni vincoli per ampliare le analisi.

L'interpretazione dei risultati mostra le difficoltà ad integrarsi della fascia giovanile di amministratori ed un alto livello di segregazione fra le amministratici femmine che hanno raggiunto i 60 anni.

. . . a Salvatore che mi ha spinto a dare il meglio, stimolandomi continuamente ed incrementando la fiducia in me stesso.

Se state leggendo queste poche parole di ringraziamento, scoprirete le battute finali di una tesi che ha richiesto molta passione e sacrificio. Per questo riservo una dedica a chi ha avuto modo di nutrirsi delle mie difficoltà, ingrassando la propria autostima.

Dedico questo traguardo ad Elisa che ha creduto in me e mi ha dato modo di sognare un futuro insieme a lei ricco di speranza.

Immatricolato bambino, ne esco uomo ed è per questo che un paragrafo speciale va alla mia famiglia ed ai miei genitori, fonte di ispirazione e motivo di orgoglio. Nonostante una carriera universitaria poco brillante, ho provato a mettermi in gioco, cercando di dare un senso ai soldi spesi ed alle loro aspettative. Nella mia mediocrità ho provato a dimostrare come si possa sopperire alla poca volontà con l'amore verso la mia famiglia che mi ha stimolato verso traguardi che sembravano inaccessibili ai miei occhi.

Se mi fossi fermato là dove il mio spirito cinico e compassato avrebbe voluto posarsi, sarei sul ciglio di un dirupo lontano da tutto e da tutti ad aspettare l'inevitabile fine.

Infine cito il *Sommo Poeta* per rafforzare la fame di sapere che dovrebbe attraversare la carne e le ossa di ogni studioso e che mi ha permesso di trovare le forze per arrivare fino a qui.

«O frati, dissi, che per cento milia
perigli siete giunti a l'occidente,
a questa tanto picciola vigilia

d'i nostri sensi ch'è del rimanente
non vogliate negar l'esperïenza,
di retro al sol, del mondo sanza gente.

Considerate la vostra semenza:
fatti non foste a viver come bruti,
ma per seguir virtute e canoscenza. »

Dante, Inferno XXVI (vv. 112-120)

Indice

Elenco delle figure

Elenco delle tabelle

Elenco degli algoritmi

Introduzione

Presentazione del problema

La storia dell'umanità è costellata di eventi che indicano chiari trattamenti non paritari attuati nei confronti di individui o gruppi di individui.

In letteratura il problema della discriminazione all'interno della società è stato spesso accostato al concetto di segregazione. I due fenomeni sono due facce della stessa medaglia, l'una ha conseguenze sull'altra e viceversa; tale binomio rappresenta uno dei più grandi ostacoli all'integrazione sociale.

In questo lavoro si mostra la definizione, la progettazione e lo sviluppo di un sistema che misura la segregazione nelle comunità che compongono una rete sociale economica.

Il fine del progetto è di scoprire se esistano valori "anomali" nelle misurazioni inerenti ad insiemi circoscritti di persone (d'ora in avanti saranno chiamati *gruppi protetti*) ed analizzare quali siano le caratteristiche di tali gruppi.

Questo studio può facilitare l'analisi di reti sociali economiche di grandi dimensioni, impossibile con le tecniche classiche dei sociologi e degli economisti prive di metodi automatici o semi-automatici che abbassano esponenzialmente i tempi di esecuzione.

In letteratura il procedimento standard è composto da diversi *step*: il primo è la definizione degli attributi sensibili che possono causare segregazione, il secondo è la specifica delle caratteristiche per delineare un profilo di un gruppo protetto, il terzo è la scelta di una misura da calcolare, il quarto è l'applicazione del calcolo della segregazione in funzione del gruppo protetto definito in precedenza ed infine vi è il commento dei risultati.

Il nostro primo passo è stata la definizione della segregazione come

"*una misura dell' isolamento di un individuo, o un gruppo di individui, dalla comunità di cui fa parte*".

Partendo da questa definizione, sono state scelte l'entropia e l'indice di isolamento, come misure di segregazione. É stata scelta la forma binaria di tali indici perchè vicina alla visione dicotomica data sopra, in cui si misura quanto un gruppo di individui si separa dal resto della comunità.

Gli indici utilizzati sono espressi in funzione di un gruppo protetto e sono stati analizzati molti gruppi protetti contemporaneamente secondo le tecniche di data mining. I risultati prodotti esprimono le misurazioni degli indici per molti gruppi protetti simultaneamente. In questo modo la fase di analisi non è circoscritta ad un unico gruppo specifico alla volta.

Tale approccio è stato utilizzato perchè a nostro parere facilita la scoperta di informazioni e l'estrazione di conoscenza dal dominio analizzato.

Di fatto, si propone un metodo di analisi alternativo tale per cui aumenta sia il carico di lavoro per quanto riguarda l'interpretazione dei risultati sia i benefici derivati da una visione più ampia e dettagliata.

Rassegna della lettura

Questa tesi propone di fornire uno studio multidisciplinare in rassegna della letteratura in merito alla segregazione, inclusi i metodi di raccolta dei dati, gli studi empirici e le tecniche per individuare possibili fenomeni discriminatori.

Per poter entrare nel contesto specifico è stato necessario documentarsi su alcune delle indagini che hanno riguardato l'analisi della segregazione ed alcuni concetti fondamentali legati ad essa.

Nel capitolo 2 si mostrano vari studi partendo dal modello dell'economista Schelling, negli anni 70, che mostra come comportamenti individuali indipendenti possano portare a fenomeni di segregazione globale.

Lo studio di Schelling viene ripreso in seguito da Roger e McKane i quali lo formalizzano in un modello matematico. In seguito si descrive la visione del sociologo Mark Fosset per cui le preferenze dei gruppi minoritari di appartenere ad un quartiere spesso destabilizzano un possibile equilibrio nella popolazione. Infine si descrivono diverse misure che quantificano la segregazione.

Come accennato in precedenza, il software creato agisce su una rete sociale economica, per questo motivo è stato necessario un breve richiamo alle reti sociali ed ad alcuni concetti ad essa collegati.

Nel capitolo 3 si riportano diversi studi sociologici che studiano le interazioni degli individui all'interno di specifiche reti sociali, in cui i nodi sono i lavoratori e gli archi sono i rapporti prefessionali che li legano.

I primi studiosi a cui ci siamo avvicinati sono Granovetter e successivamente Montgomery che sottolineano l'importanza della forza dei legami nelle reti sociali ed illustrano come possa essere determinante per l'insorgere di comportamenti discriminatori.

In seguito si descrive il punto di vista di Van der Leij, per cui la segregazione è identificata come una forma di discriminazione.

Si riporta successivamente il modello economico di Armengol e Jackson che mostra quanto una rete sociale ed i suoi legami possano influenzare il tasso di occupazione all'interno di una comunità. Infine si descrive il feno-

meno chiamato "*omofilia*" per cui all'interno di una rete sociale tendono a formarsi gruppi di individui omogenei sotto il punto di vista demografico e culturale.

Contenuto della tesi

La tesi verte sulla definizione, la progettazione e lo sviluppo di un sistema per la misura della segregazione in funzione di un gruppo protetto all'interno di una rete sociale economica, in cui i nodi sono le aziende e gli archi sono il numero di amministratori che possiedono in comune.

Gli indici utilizzati sono stati l'entropia e l'indice di isolamento perchè conformi alla definizione di segregazione che si è definita.

Nel capitolo 4 si descrive il processo di data mining. Per prima cosa si giustifica il motivo della tesi, in seguito sono esposti i vari step tipici dello sviluppo di un software. Si illustra quali funzionalità il sistema deve avere nella sezione "analisi dei requisiti". Si specifica cosa deve misurare e in che modo nella sezione "definizione della segregazione". L'argomento trattato nella sezione "definizione della rete sociale di analisi" è il tipo di rete che il sistema deve analizzare mentre il dettaglio delle istanze utilizzate è affrontato in "descrizione delle rete sociali di analisi".

Nel capitolo 5 si mostra la fase implementativa con la descrizione delle varie funzionalità implementate in pseudo codice.

Il capitolo 6 mostra alcuni test effettuati su una rete sociale economica di aziende nel contesto informatico ed una nel contesto finanziario. I risultati ottenuti vengono descritti ed analizzati.

Lo scopo del progetto è capire se esistano *gruppi protetti*, ovvero insiemi di individui con caratteristiche socio-culturali comuni, che presentano anomalie nei valori degli indici utilizzati e per quali tipi di soggetto tali forme siano massimizzate.

Il software non produrrà un valore degli indici in funzione di un unico gruppo protetto bensì fornirà una rappresentazione degli indici in funzione dei gruppi protetti individuati. Tale forma è stata scelta per mostrare uno scenario più espressivo e meno "*piatto*". Da un certo punto di vista è come se il software risentisse in maniera minore della definizione dei gruppi protetti e quindi dei pregiudizi del suo creatore.

Secondo il nostro parere, mostrare i risultati per più gruppi protetti contemporaneamente offre una visione più emancipata della realtà.

Affinchè i sistemi decisionali siano emancipati dagli stessi sviluppatori, bisogna in primo luogo pensare da emancipati, poiché i nostri programmi sono influenzati direttamente dal nostro modo di pensare.

Capitolo 1

Segregazione

La segregazione ha una storia antica nella società americana. Fra gli anni '40 e '50 furono proposti i primi indici di segregazione nella "*Rivista Sociologica Americana*". I maggiori studi metodologici sono stati effettuati nella misura della segregazione fra due gruppi di popolazioni, tipicamente bianchi e neri o maschi e femmine. Duncan [6] afferma come le misure di segregazione siano costruite da ingenue nozioni piuttosto che progettate attraverso concettualizzazioni pertinenti.

> [Segregation] "*is a concept rich in theoretical suggestiveness and of unquestionable heuristic value. Clearly we would not wish to sacrifice the capital of theorization and observation already invested in the concept. Yet this is what is involved in the solution offered by naive operationalism, in more or less arbitrarily matching some convenient numerical procedure with the verbal concept of segregation ...*"

Nonostante la sua visione, fino agli anni '80 non fù mai proposto un approccio teorico completo per la misurazione della segregazione. James e Tauber nel 1985 [13] dimostrarono come alcuni indici, in precedenza utilizzati come misure di disugualianza, potessero essere utilizzati anche per misurare per la segregazione.

Reardon e Firebaugh nel 2002 hanno illustrato come la società si sia trasformata grazie alla globalizzazione e come sia più eterogenea rispetto ai primi studi in cui si prendevano in considerazione misure di segregazione in merito a due gruppi (bianchi e neri). Tali strumenti risultano inadeguati per descrivere i pattern complessi, e per ciò nel loro studio introducono indici "multi gruppo", come estensione dei primi indici.

1.1 Modello di Schelling

L'economista americano Schelling ha ideato fra il 1969 ed il 1971 alcuni modelli di segregazione [27] in cui ha dimostrato che il fenomeno dell'omofilia

(ad esempio il formarsi di comunità omogenee dello stesso colore) totale o parziale (accettazione di una comunità con una proporzione di invididui indesiderati entro una certa soglia) può portare alla totale segregazione.

Schelling affronta la segregazione come "*consapevolezza, conscia o inconscia,*" che influenza le decisioni.

Attraverso il suo studio suggerisce come piccole variazioni in preferenze o scelte individuali non casuali possono portare alla formazione di comunità omogenee fra loro. Sviluppa un modello in cui dimostra come lievi differenze nelle preferenze da parte di un individuo x di risiedere in un posto in cui vi è una maggioranza di individui simili ad egli, possa portare a risultati in cui interi quartieri sono in prevalenza formati da individui di etnia simile.

I comportamenti individuali indipendenti l'uno dall'altro, di individui con lievi differenze di razza possono portare al formarsi di uno scenario in cui intere aree geografiche sono in prevalenza occupate da bianchi ed in minoranza da neri, o viceversa.

Un primo modello analizzato è una simulazione in cui alcuni individui appartenenti a due gruppi distinti che distribuiscono se stessi in quartieri, mentre un secondo modello tratta gli spazi a compartimenti. Schelling spiega come la segregazione sia misurata in funzione di vari attributi come il sesso, età, lingua, stipendio, religione, colore ed altre ancora; viene osservata come risultato di prassi di organizzazioni, alcune deliberatamente organizzate, oppure dall'interazione fra persone che attuano scelte individuali discriminatorie. Alcuni tipi di segregazione sono logiche conseguenze di altri tipi: la residenza sicuramente è correlata alla posizione del proprio lavoro e del trasporto, per esempio gli appartamenti all'interno di un condominio saranno raggruppati per il loro valore e quindi per l'affitto necessario e questo porta ad avvicinare le persone che possiedono stipendi simili [22].

Schelling sottolinea come sia importante capire i meccanismi della segregazione economica (scelte individuali oppure organizzate o indotte economicamente) per risolvere i problemi di *equità sociale* e comprendere quelli della segregazione organizzata aiuta a difendere i *diritti civili* dell'individuo. Appare evidente come la segregazione indotta economicamente sia in correlazione con la discriminazione. Inoltre se tale processo viene portato avanti, l'alienazione, l'ostilità, la semplice abitudine, la paura possono portare alla tendenza all'allontanamento.

Vincoli quantitativi

Schelling esegue i suoi esperimenti manipolando la distribuzione delle preferenze dei gruppi e la distribuzione dell'etnie, questo poiché una analisi quantitativa di un fenomeno offre un quadro descrittivo della realtà presente e possiede pochi vincoli logici. Vincoli di questo tipo spingono un individuo a scegliere se accettare o meno un college se il 10% degli studenti possiede determinate caratteristiche (ad esempio è nero), o a comprare una casa in

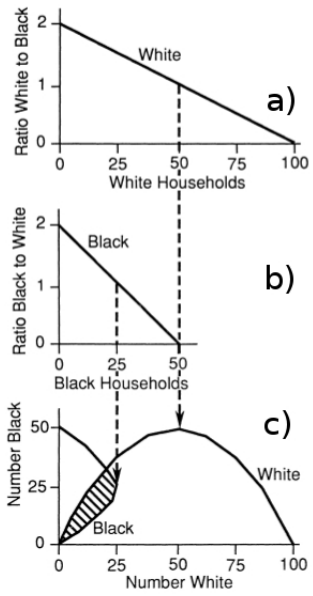


Figura 1.1: Vincoli quantitativi nel modello di Schelling, da [5]

un quartiere se tre vicini su cinque sono della stessa religione o etnia e così dicendo.

Quando ogni individuo di una comunità insiste sulla necessità di risiedere in un quartiere dove è presente una maggioranza locale della sua etnia, esiste solo una soluzione: la *segregazione totale*, in questo caso si parla di *equilibrio stabile* poiché nessun individuo si muoverà in un'area dominata da un gruppo a cui egli non appartiene.

Un tipico schema utilizzato prevedeva lo studio di cento individui bianchi e cinquanta di colore, con identiche distribuzioni di preferenza/tolleranza: nella figura 1.1 al punto a si osserva una pianificazione rettilinea degli individui ovvero un grafico in cui l'asse orizzontale delle x rappresenta il numero α di bianchi (neri se si osserva il punto b) ordinati in base alla tolleranza verso i neri (bianchi se si osserva la figura al punto b): i primi individui bianchi a partire da sinistra sono quelli che accettano di convivere in quartieri in cui la proporzione di neri su bianchi è di due a uno. Sull'asse delle y viene rappresentato il limite superiore della tolleranza accettata ovvero nel caso dei bianchi il rapporto $\frac{neri}{\alpha}$. Nel punto c della figura sull'asse delle y è riportato il grafico del punto b con la sola eccezione che la tolleranza accettata non è espressa in proporzione ma in valore assoluto ovvero il numero di individui accettati dell'altra razza, parimenti sull'asse delle x è riportato il grafico del punto a con la stessa trasformazione della tolleranza in valore assoluto. L'area di intersezione delle due parabole rappresenta una combinazione valida di bianchi e neri, il luogo in cui entrambe le razze coinvivono accettando l'una la presenza dell'altra.

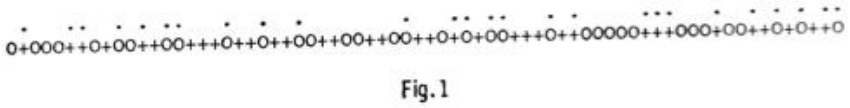


Figura 1.2: Risultati del modello di prossimità spaziale di Schelling

Modello spaziale

É un modello che sperimenta la situazione in cui due etnie, i bianchi ed i neri, devono decidere come distribuirsi in un'area o su linea, in accordo alle preferenze dei vicini, tenendo in considerazione prevalentemente la loro razza. Come si vede dai risultati originali riportati in seguito un esperimento reale nella figura 1.2, i bianchi ed i neri etichettati da un cerchietto e da una croce sono rappresentati su di una linea: si notano formazioni di piccoli gruppi omogenei, mentre i cerchi e le croci con il puntino sopra sono gli individui insoddisfatti della loro posizione. Un altro esperimento simile è stato effettuato rendendo lo spazio a compartimenti e quindi dando la possibilità di essere all'interno o all'esterno di un quartiere, il tutto cercando di rispettare le preferenze individuali ed eventuali esigenze di tipo numerico (vincoli sulla frazione di vicini dello stesso tipo). Nel modello spaziale i possibili parametri da variare sono la grandezza del quartiere, i vincoli sulla frazione di vicini dello stesso tipo, la percentuale dei tipi, le regole che governano i movimenti e la configurazione originale ed a seconda della loro configurazione la segregazione può diventare più o meno estrema. Schelling ha notato che diminuendo la percentuale di individui appartenenti ad una minoranza, il numero di cluster diminuisce in modo più che proporzionale a causa dei vincoli di preferenza che costringono gli individui a raggrupparsi in comunità sempre più unite.

Modello a quartiere limitato

In questo tipo di esperimento viene cambiata la definizione di quartiere, infatti non è più ogni singolo individuo che lo definisce attraverso la sua stessa posizione, ma viene offerta una definizione comune di quartiere e di confini. In questa ipotesi possono essere inclusi esempi come la partecipazione ad un lavoro, o in un ufficio, in un ristorante o in un ospedale. In questo modello ogni individuo ha la preferenza per un quartiere, e può accettare di trasferirsi all'interno di uno composto da individui di tipo differente entro certi vincoli di frequenza o scegliere di andare in qualche altro posto. La tolleranza diventa un concetto relativo, concepito localmente per ogni singola zona, infatti ogni individui possiede una preferenza e per ogni quartiere esiste un insieme di persone con una soglia minima di tolleranza ma chiaramente le soglie non saranno tutte uguali, ogni quartiere avrà la sua soglia minima di tolleranza.

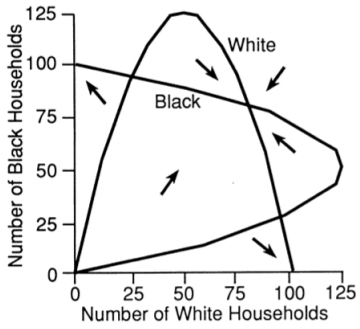


Figura 1.3: Equilibrio nel modello a quartiere limitato di Schelling, da [5]

L'informazione è perfetta ovvero tutti sanno le frequenze degli individui di entrambi i gruppi dentro ogni area al momento in cui scelgono ma non possono prevedere le mosse degli altri individui, ne eventuali spostamenti futuri. L'equilibrio viene trovato solamente in condizioni in cui un quartiere è predominato da un tipo di individui (tutti neri oppure tutti bianchi), quale sia viene determinato in fase iniziale a seconda di come sia la combinazione iniziale e la velocità di occupazione del quartiere, inoltre è importante anche la distribuzione di tolleranza, più alta è la tolleranza generale, maggiore sarà la possibilità di arrivare ad una situazione in cui le due razze sono integrate. Si veda la figura 1.3, in cui le due campane rappresentano la curva di distribuzione degli individui in funzione della tolleranza che hanno nei confronti dell'altra razza. La loro intersezione rappresenta visivamente quanto sia grande l'area in cui i due tipi di individui sono integrati.

Risultati secondo Schelling

Gli studi di Schelling furono condotti ponendo l'ipotesi che la tolleranza fosse una *variabile individuale*, influenzata dall'unico fattore preso in considerazione (nel caso di Schelling si concentra sui colori della pelle bianco e nero). Entrambi i modelli studiano i limiti di tolleranza che gli individui possono avere e simulano vari possibili scenari di integrazione dal punto di vista spaziale (ovvero del luogo di residenza). Questi modelli dimostrano la loro limitatezza proprio perchè non permettono comportamenti speculativi, scostamenti rilevanti, azioni organizzate o errate percezioni. I modelli possono comunque essere utilizzati per descrivere il fenomeno di nuove minoranze che entrano in gioco e scoinvolgono la situazione di equilibrio.

1.2 Paradosso delle preferenze minoritarie deboli

Lo studioso Mark Fosset [10] attraverso un modello ABM[1] ha sfidato la visione prevalente che imponeva la segregazione come conseguenza di azioni collettive deliberate e che operano con l'intenzione di aggirare le leggi vigenti in materia discriminatoria. Il modello da lui proposto mostra come le preferenze delle minoranze sono sufficienti per la segregazione in assenza di pratiche discriminatorie.

Lo stesso scienziato sottolinea il suo approccio, non traendo conclusioni sui modelli di segregazione per le città ma indagando nel profondo della questione teorica:
"*Can ethnic preferences and social distance dynamics create and sustain significant levels of ethnic segregation in a theoretical system where discrimination is absent?*" (cit. [10]).
Egli afferma attraverso un'espressione chiamata "*paradosso delle preferenze minoritarie deboli*" che i gruppi minoritari spesso destabilizzano un possibile equilibrio, infatti le loro preferenze di appartenere ad un quartiere spesso eccedono la frequenza di distribuzione del gruppo stesso nella popolazione e ciò porta a una rottura dell'integrazione: il problema nasce quando una minoranza con il 10% di frequenza aspira ad avere un quartiere con una frequenza di individui dello stesso tipo superiore a quella assoluta. Fosset afferma che l'integrazione consista nel ritrovare all'interno dei quartieri una distribuzione degli individui che rispetti la distribuzione globale della popolazione: se in una nazione esiste il 90% di bianchi ed il 10% di neri, all'interno di una città affinchè vi sia integrazione queste proporzioni devono essere rispettate.

Più piccola è la dimensione della minoranza maggiori sono gli effetti di segregazione che si possono verificare: le sue preferenze incideranno notevolmente sulla formazione di piccoli gruppi con distribuzioni simili a quelle da loro desiderate e non permetteranno alcuna integrazione.

1.3 Il quadro unificato di Roger e McKane

Questi due scienziati hanno ripreso il modello di Schelling e lo hanno generalizzato, traducendolo in un modello matematico: le razze, i quartieri e la tolleranza sono stati sviluppati come variabili e funzioni. In primo luogo nel loro lavoro è stato sottolineato come il concetto di segregazione sia per sua natura spaziale. La distanza in Schelling veniva vista come geografica, anche se potrebbe essere vista in termini di distanza di rapporti interpersonali.

Identificano quattro componenti principali che specificano il modello di Schelling:

- la rete

[1] agent-based model è una classe di modello computazionale per simulare le azioni e le interazioni di agenti autonomi

- le condizioni iniziali
- la funzione di soddisfazione
- le probabilità di trasferimento.

Matematicamente il modello è facilmente formalizzato attraverso l'utilizzo di una rete, in cui esistono N siti e degli archi fra i siti se sono fra loro "*vicini*": il quartiere di un nodo i è definito con il simbolo δ_i ed è l'insieme di tutti i nodi vicini di i. Per semplicità non sono considerate reti con archi multipli fra due nodi e senza archi che puntano da un sito a se stesso. In ogni momento un sito può essere occupato da al massimo un agente o può essere disponibile, σ_i simula lo stato del sito e può essere espresso come

$$\sigma_i \begin{cases} 1 & \text{Se il sito } i \text{ è occupato da un agente di tipo A} \\ -1 & \text{Se il sito } i \text{ è occupato da un agente di tipo B} \\ 0 & \text{Se il sito } i \text{ è è disponibile} \end{cases}$$

in questo modo lo stato dell'intero sistema ad un tempo specifico, corrispondente alla posizione di ogni agente nella rete, è dato dal vettore $\sigma = (\sigma_1, \ldots, \sigma_N)$, per ogni coppia i,j di siti:

$$\sigma_i\sigma_j \begin{cases} 1 & \text{Se } i \text{ e } j \text{ sono occupati da agenti dello stesso tipo} \\ -1 & \text{Se } i \text{ e } j \text{ sono occupati da agenti di tipo differente} \\ 0 & \text{Se almeno uno sito è disponibile} \end{cases}$$

Da qui si vede come il soggetto principale di questo modello siano i siti piuttosto degli agenti e di quanto sia rilevante quali siti sono occupati.

La frazione dei siti disponibili è definita come $\rho = \frac{1}{N}\sum_i(1-|\sigma_i|)$, lo stato iniziale è il vettore σ al tempo zero, implementato in modo casuale.

Per descrivere i motivi per cui un agente si sposta da un sito ad un altro, viene implementata la funzione di soddisfazione che assume valori più bassi quando i vicini sono prevalentemente di tipo opposto. Viene introdotto così un vettore $s = (s_1, \ldots, s_n)$ in cui la variabile s_i è compresa nell'intervallo [0;1] e codifica la soddisfazione di un agente che occupa il sito i (se i è disponibile è zero). Non viene fatta alcuna restrizione sulle specifiche che la funzione deve avere ma viene specificato che dipende solamente dal numero di vicini simili o diversi che un agente possiede, matematicamente significa che σ_i è funzione dei numeri $\sigma_i\sigma_j$ per $j \in \delta_i$.

La funzione di soddisfazione s_i di un sito i occupato può essere espressa anche in funzione della frazione di vicini che occupano il sito e che contengono agenti di tipo opposto, questa funzione viene chiamata x_i dove $0 \leq x_i \leq 1$, e $x_i = 0$ se il sito i è disponibile oppure è circondato da aree disponibili.

In caso contrario viene calcolato come

$$x_i = \frac{\sum_{i\in\delta_i}(|\sigma_i\sigma_j| - \sigma_i\sigma_j)}{2 \cdot \sum_{i\in\delta_i}(|\sigma_i\sigma_j|)} \tag{1.1}$$

Viene rappresentato con $\sigma^{(ij)}$ la situazione in cui lo stato della rete e la soddisfazione annessa cambiano in seguito ad uno scambio di posti fra i e j e per determinare quale scambio deve essere effettuato in situazioni in cui gli agenti sono disposti a spostarsi viene definita la probabilità di trasferimento $T_{ij}(\sigma)$, fornendo la possibilità che i siti verranno selezionati ed il loro contenuto scambiato. Di seguito viene mostrato un vincolo sulla somma di tutte le probabilità di trasferimento fra agenti che deve risultare uguale ad uno (caratteristica basilare delle probabilità), e deve essere soddisfatto da tutte le σ:

$$\sum_{i,j} T_{ij}(\sigma) = 1$$

Con $P(\sigma, t)$ si scrive la probabilità che il sistema sia nello stato σ al tempo t, le condizioni iniziali sono specificate come $P(\sigma, 0)$. Matematicamente l'evoluzione del sistema è espressa come una somma di tutti i possibili trasferimenti fra gli agenti, che a sua volta sono in funzione dello stato dei nodi e della probabilità che questi siano in quel determinato stato al tempo t.

$$P(\sigma, t+1) = \sum_{i,j} T_{ij}(\sigma^{(ij)} \cdot P(\sigma^{(ij)}, t))$$

Poiché è necessario imporre dei vincoli su T, la soluzione adottata dai due studiosi specifica che T_{ij} deve essere presa come il prodotto di tre componenti:

(i) la probabilità di selezionare un agente nel sito i per avere la possibilità di muoversi (zero se i è disponibile e non crescente in s_i).

(ii) la probabilità di selezionare un sito j come destinazione dello spostamento (non incrementa nella distanza da i a j)

(iii) una misura di quanto sia desiderabile il sito j per un agente i (non decresce in $s_{(ij)}$).

Mentre in Schelling le informazioni quantitative vengono mostrate attraverso grafici, per Roger e McKane, questo metodo appare utile solamente in reti strutturate come reticoli quadrati ed è quindi necessario introdurre statistiche numeriche appropriate che catturino alcune caratteristiche del vettore σ. Ne esistono di due categorie: quelle che contano alcune configurazioni locali di agenti e quelle che osservano lo stato globale del sistema. Del primo tipo, è analizzata la densità che è una buona rappresentazione delle statistiche locali ed è definita come

$$x = \frac{\text{numero di archi fra agenti di tipo opposto}}{\text{numero di archi fra agenti di ogni tipo}}$$

Da notare che x è la media su tutta la rete della quantità locale x_i introdotta nell'equazione 1.1. Del secondo tipo, molti modelli misurano la grandezza delle regioni che contengono agenti di un solo tipo, questo comportamento noto come "*clustering*" produce da analizzare sia le dimensioni dei cluster che le loro forme, anche se questo tipo di misura porta ad un difficile processo di apprendimento della distribuzione della dimensione dei cluster.

1.4 Misure della segregazione

1.4.1 Basate sull'iterazioni sociali

Modello di Freeman

Nel 1978 lo studioso Freeman introduce un metodo per misurare la segregazione vista come una limitazione di interazioni fra le persone, legata a determinate caratteristiche degli individui coinvolti.

É definita come un pregiudizio che influenza le interazioni altrimenti regolate come una rete casuale di rapporti interpersonali. Osservando la rete sociale come un grafo diretto consistente come un insieme di punti $A = \{a_i\}$ ed un insieme di archi (o coppie non ordinate) $R = \{r_k\} = \{(a_i, a_j)\}$ in cui i punti rappresentano gli individui e gli archi i rapporti interpersonali.

Viene definita una classe di di persone, tramite qualche criterio, che possono essere soggette a limitazioni nelle loro interazioni e tale insieme viene definito come A_g. Siano

$$\begin{aligned} m &= \text{il numero di punti di } A, \\ m_g &= \text{il numero di punti di } A_g, \\ n &= \text{il numero di punti di } R \end{aligned}$$

ed e_k^* una variabile casuale definita come

$$e_k^* = \begin{cases} 1 & \text{Se e solo se } r_k \text{ è un arco fra un individuo} \in A_g \text{ed uno} \notin A_g \\ 0 & \text{altrimenti} \end{cases}$$

Allora $e^* = \sum_{k=1}^{n} e_k^*$ è il totale di archi che collegano i punti di A_g con gli individui esterni, questa è la misura generale della segregazione.

Più e^* è grande più legami inter-classi ci sono e quindi sono sistematicamente integrate. Ovviamente non è una misura plausibile poiché dipende dal numero totale di persone coinvolte e sulla proporzione di chi cade nella sottoclasse designata. Per questo viene trasformata grazie al confronto di reti con relazioni generate in modo casuale in modo tale da poter confrontare una rete sociale reale con un "*prototipo*" in cui oggettivamente manca la segregazione. A tale fine è definita la quantità s come la differenza degli archi "inter-classe" fra il modello reale e quello in cui gli archi sono generati casualmente.

$$s = max\{E(e^*) - e^*, 0\}$$

Questa misura è la misura naturale della segregazione ma finché viene espressa in funzione del numero di archi e di nodi della rete non permetterà comparazioni fra grafi di dimensioni differenti.

É utile quindi normalizzarla con il massimo valore ottenuto dal modello generato casualmente. Tale quantità viene espressa come S, misura di segregazione che può variare fra zero (massima integrazione) ed uno (segregazione completa).

$$S = \frac{s}{E(e^*)}$$

Il valore atteso $E(e^*)$ di cui parla Freeman è ottenuto attraverso una serie di passaggi esposti di seguito. Dati m individui, il numero totale di archi è

$$N = \binom{m}{2} = \frac{m \cdot (m-1)}{2}$$

ed il totale possibile numero di archi "inter-classe" è $x = m_g(m - m_g)$, con la generazione degli archi in modo casuale, la probabilità che un arco sia "inter-classe" è definita come

$$p = \frac{x}{N}$$

in modo complementare la probabilità che un arco colleghi due individui all'interno della stessa classe è definita come

$$q = 1 - p$$

A questo punto, dato un numero di archi z, il valore atteso degli archi "inter-classe" è definito come

$$E(e^*) = z \cdot p$$

Indice spettrale di segregazione - SSI

Federico Echenique sviluppa un indice di segregazione che disaggrega a livello degli individui e che da una misura di segregazione proporzionale al livello di segregazione degli agenti con cui interagisce.

É un indice che si basa sulle *interazioni sociali* come ad esempio fra bianchi e bianchi o fra neri e neri ed è stato applicato per misurare la segregazione residenziale ed a scuola. Echenique afferma come tutti gli indici che vogliano misurare la segregazione debbano possedere tre proprietà fondamentali

- *Monotonicità*: se tutti gli individui di una rete A posseggono una quota maggiore di interazioni con agenti dello stesso tipo, rispetto ad una rete B, allora A è più segregata di B.

- *Linearità*: ogni individuo è segregato in maniera linearmente proporzionale agli agenti con cui interagisce.

- *Omogeneità*: se tutti gli individui di una rete hanno metà delle loro interazioni con individui dello stesso tipo allora l'indice di segregazione deve esprimere questa situazione e deve essere un mezzo (in modo da normalizzare l'indice).

Nel suo studio spiega come un indice chiamato "*Indice Spettrale di Segregazione*, con l'acronimo di SSI, sia in grado di rispettare tutte le proprietà

sopra elencate. Si intende con *primo ordine di segregazione* la quota di interazioni sociali che si hanno con gli individui del proprio gruppo, mentre il *secondo ordine* rappresenta la media di tutte le interazioni, nel proprio gruppo sociale, del primo grado di segregazione. Il *SSI* di un individuo è il limite per $n \to \infty$ dell'n^{esimo} ordine di segregazione dell'individuo.

SSI offre vantaggi rispetto alle altre misure esistenti di segregazione come l'invarianza rispetto alle partizioni di una città, inoltre permette di studiare come sono segregati gruppi multipli di minoranza e di confrontarli fra loro.

Possiede la capacità di misurare la segregazione a livelli disaggregati, permettendo di misurare l'intensità di un cluster per razza o svelare i blocchi di città più segregati: cattura le interazioni inter-razziali grazie alla possibilità di pesare l'intensità dei rapporti fra la stessa razza.

Un punto a sfavore risiede nella qualità delle informazioni che si possono ottenere sulle interazioni sociali.

Processo di calcolo del SSI Dato un insieme di individui V e le informazioni sulla interazione (possibile, concreta e sulla sua quantificazione) di due individui qualsiasi, si presuppone che per ogni coppia di individui si conosca se loro interagiscono o meno ed in che quantità. Per ogni coppia di individui v e v', $r_{vv'} \geq 0$ rappresenta la natura della loro relazione. Se è strettamente maggiore di zero allora i due individui hanno una relazione, altrimenti no.

Le informazioni sulle interazioni possono essere espresse in una matrice R, di dimensione $V \times V$, in cui ogni cella è definita da $r_{vv'}$ (pensato come una frazione del tempo che v spende con v'). Viene espresso un vincolo di bilancio per le interazioni di un singolo individuo

$$\sum_{v' \in V} r_{vv'} = 1 \qquad \forall v \in V$$

Inoltre viene assunto che se $r_{vv'} = 0$ allora anche $r_{v'v}$ è uguale a zero, nel caso in cui entrambe le quantità sono maggiori di zero allora possono assumere valori diversi, questo perchè una relazione può avere livelli differenti di importanza a seconda del soggetto e delle sue interazioni globali.

L'insieme V di tutti gli individui viene filtrato per una razza h che è fissata e viene costruita la matrice B da quella originaria R filtrando tutte e sole le relazioni che hanno entrambi gli individui della razza h.

La matrice B indica quindi le interazioni fra gli individui della stessa razza h e si può ipotizzare che il valore $r_{vv'}$ sia uguale a zero se gli individui non sono vicini oppure uguale all'inverso dei vicini di v se ogni interazione assume la stessa importanza. Con questo tipo di matrice è possibile analizzare l'intensità dei contatti fra razze diverse poiché nella matrice B se un individuo v interagisce solo con un individuo v' della razza h allora $r_{vv'} = 1$ mentre se v interagisce con 9 membri di un'altra razza oltre ad interagire con v' allora $r_{vv'} = \frac{1}{10}$ e sarà l'unico elemento non nullo della riga v.

L'indice *SSI* rappresenta il più grande autovalore della corrispondente sottomatrice irriducibile di B. I singoli *SSI* sono ottenuti distribuendo le componenti *SSI* fra gli individui che utilizzano l'autovettore corrispondente al più grande autovalore.

1.4.2 Indici dicotomici basati sulla frequenza

Indice di difformità [2]

Sviluppato da Jahn, Schmid, and Schrag nel 1947 rappresenta una misura di uniformità, supponendo che la città sia divisa in N sezioni, questo indice misura per ogni sezione la percentuale della popolazione di un gruppo che deve cambiare sezione per riequilibrare le proporzioni secondo le percentuali globali.

$$\text{indice di difformità} = \frac{1}{2} \cdot \sum_{i=1}^{N} \left| \frac{neri_i}{neri_{totali}} - \frac{nonneri_i}{nonneri_{totali}} \right| \tag{1.2}$$

dove $neri_i$ è il numero di individui di colore nell'area i e $neri_{totali}$ è il totale di individui di colore nell'intera città ed in modo reciproco per gli individui bianchi.

Indice di isolamento[2]

Come evidenzia lo studioso Blau nel 1977 i neri possono essere distribuiti equamente nelle aree residenziali della città ma in realtà rappresentano una scarsa esposizione ai non-neri se sono una percentuale relativamente elevata della città. L'indice di isolamento misura come i neri siano esposti l'uno all'altro piuttosto che ai non-neri, è calcolato come la media ponderata della minoranza di ogni popolazione minoritaria.

$$\text{indice di isolamento} = \sum_{i} \left(\frac{neri_i}{neri_{totali}} \cdot \frac{neri_i}{persone_i} \right) \tag{1.3}$$

dove $persone_i$ si riferisce al totale degli individui nell'area i.

L'indice è definito nell'intervallo [0;1], raggiunge il suo massimo valore quando tutti gli individui di un gruppo protetto sono isolati, ovvero sono raggruppati in una unica comunità, viceversa raggiunge il minimo quando tali individui sono equamente distribuiti fra tutte le comunità mantenendo le proporzioni globali: in un esempio in cui esistono cinque comunità con una popolazione complessiva di mille individui, composta da un gruppo protetto di individui di colore del 20 %, l'indice assumerà il suo valore minimo

[2] Possiede due proprietà indesiderabili: dipende esplicitamente dal modo arbitrario in cui le città sono partizionate in sezioni ed in secondo luogo non è definita quando si cerca di valutare la segregazione a livello degli individui.

quando, tutti gli individui sono equamente distribuitivi fra tutte le comunità, rispettando la proporzione del 20%, in tale maniera l'isolamento sarà minimo poiché tale gruppo protetto sarà esposto agli altri individui rispettando le proporzioni complessive.

1.4.3 Indici multigruppo

Reardon e Firebaugh nel 2002, hanno illustrato come la società si sia trasformata in una realtà formata da molteplici comunità eterogenee, più complessa rispetto ai primi studi in cui si prendevano in considerazione misure di segregazione in merito a due gruppi (tipicamente bianchi e neri). Tali strumenti risultano inadeguati per descrivere i pattern complessi, per ciò introducono indici "multi gruppo".

Viene utilizzata la seguente notazione: t indica la dimensione e π indica la proporzione, gli indici i e j indicano le unità organizzative (scuole), e gli indici m e n indicano i gruppi (ad esempio gruppi raziali). Da qui

$t_j =$ il numero di invididui nell'unità organizzativa j
$T =$ il numero totale di casi, notare che $\sum_j \frac{t_j}{T} = 1$
$\pi_m =$ la proporzione del gruppo m (ad esempio la proporzione di neri)
$\pi_{jm} =$ la proporzione del gruppo m dell'unità j (i neri nella scuola j).

Poiché le misure di segregazione sono viste come funzioni delle proporzioni dei gruppi (π_m, π_{jm}) spiccano in primo piano due misure della variazione del gruppo di appartenenza:

$$I = \sum_{m=1}^{M} \pi_m \cdot (1 - \pi_m) \tag{1.4}$$

$$E = \sum_{m=1}^{M} \pi_m \cdot \ln\left(\frac{1}{\pi_m}\right) \tag{1.5}$$

Il primo rappresenta l'indice *Simpson's interaction* ed il secondo rappresenta un caso particolare di indice dell'entropia, sviluppato da Theil.

Entrambi rappresentano misure della diversità della popolazione, ed entrambi sono uguali a zero se e solo se tutti i membri della popolazione appartengono ad un unico gruppo e tutti e due, raggiungono il massimo quando tutti gli individui sono uniformemente distribuiti fra gli M gruppi ($\pi_m = 1/M$ per tutti gli m).

James and Taeuber [13] hanno indicato quattro principali criteri di valutazione che specificano come gli indici dovrebbero rispondere ai cambiamenti nella distribuzione dei gruppi fra le unità organizzative, di seguito elencati.

- **Equivalenza organizzativa** se una unità organizzativa è divisa in k unità, ognuna con le stesse proporzioni nei gruppi dell'unità originale,

la segregazione rimane invariata. Equivalentemente se k unità organizzative con le stesse proporzioni dei gruppi sono fuse in un'unica, la segregazione rimarrà immutata.

- **Invarianza nella dimensione** se il numero di persone di ogni gruppo m in ogni unità organizzativa j è moltiplicato per un fattore costante p, l'indice di segregazione rimane invariato.

- **Principio dei trasferimenti**[3] se un individuo di un gruppo m è spostato da una unità organizzativa i ad una j, dove la proporzione del gruppo m è maggiore nell'unità i che in j ($\pi_{im} > \pi_{jm}$), allora la segregazione è ridotta. Il rispetto di tale principio è fondamentale per un indice di segregazione perché offre la possibilità di registrare la variazione di segregazione a seconda degli spostamenti di un individuo.

- **Invarianza nella composizione** se il numero di persone del gruppo m in ogni unità viene incrementato di un fattore costante p, ma il numero e la distribuzione di persone di tutti gli altri gruppi rimane invariata, allora parimenti la segregazione resta immutata.

Nonostante nel criterio dei trasferimenti siano inclusi i due modi di trasferimento, può accadere che gli indici di segregazione multi gruppo si comportino in maniera differente a seconda che si parli di scambio o di trasferimento di individui fra gruppi ecco perchè viene introdotto un quinto criterio.

- **Principio di scambio** se un individuo di un gruppo m in un'unità organizzativa i si scambia con un individuo di un gruppo n di una unità organizzativa j, dove la proporzione di persone del gruppo m è maggiore nell'unità i che in j ($\pi_{im} > \pi_{jm}$), e la proporzione del gruppo n è maggiore nell'unità j che in i ($\pi_{jn} > \pi_{in}$), la segregazione è ridotta.

Reardon e Firebaugh hanno aggiunto ai criteri sopra elencati, due proprietà di scomposizione che sono desiderabili per gli indici di segregazione.

- **Scomponibilità additiva dell'unità organizzativa** se J unità organizzative sono clusterizzate in K cluster, allora la misura di segregazione dovrebbe essere decomposta in una somma di componenti indipendenti all'interno e fra cluster differenti.

- **Scomponibilità additiva del gruppo** se M gruppi sono clusterizzati in N super gruppi, allora la misura di segregazione dovrebbe essere decomposta in una somma di componenti indipendenti all'interno e fra super gruppi differenti.

[3] Vengono inclusi nello stesso criterio due modi di trasferimento che si comportano nello stesso modo: il "*one way*" e quello di scambio chiamato "*two way*"

Questi tipi di proprietà sebbene non siano necessarie per la definizione di una significativa misura di segregazione, tornano utili in analisi pratiche in cui le unità organizzative ed i gruppi vengono clusterizzati.

Approcci per il calcolo di una misura multi gruppo

Una misura di segregazione multi gruppo può essere calcolata in diversi modi, Reardon e Firebaugh descrivono quattro approcci differenti, che riflettono modi alternativi di concepire la segregazione:

1. una funzione che misura la sproporzione fra le proporzioni dei gruppi riguardo alle unità organizzative, questa concettualizzazione collega la misura di segregazione a quella di disuguaglianza.

2. una associazione fra gruppi ed unità organizzative, così la segregazione viene associata a misure di associazione come il χ^2 o G^2.

3. una variazione della diversità delle unità (ad esempio come una variazione della diversità etnica nelle scuole), in primo luogo viene definita una misura della diversità totale della popolazione, ed in seguito viene definita la segregazione come percentuale di questa diversità totale calcolata con le differenze inter unità (ad esempio da scuola a scuola) nelle proporzioni del gruppo.

4. una misura di segregazione multi gruppo come una media pesata di indici di segregazione dicotomici.

1.5 Modello strutturale di segregazione di Mele

Angelo Mele [16] sviluppa un modello dinamico di formazione strategica della rete e dimostra l'esistenza di un punto di equilibrio unico e stazionario, che caratterizza la probabilità di trovare una specifica rete nei dati, mentre i modelli trattati in passato avevano equilibri multipli. Come conseguenza, i parametri strutturali possono essere stimati usando una sola osservazione della rete in un singolo punto nel tempo. La stima è difficile perché la valutazione esatta della probabilità richiede costi di calcolo elevati. Per aggirare questo problema, egli propone una catena di Markov attraverso un algoritmo bayesiano che evita la valutazione diretta della probabilità. Questo metodo riduce drasticamente l'onere computazionale di stimare la distribuzione a posteriori e consente inferenza in grossi modelli dimensionali.

Il modello viene utilizzato per studiare come alcune differenti politiche per eliminare la segregazione possano influenzare la struttura della rete in equilibrio. I modelli strategici interpretano la rete come l'equilibrio risultante da un gioco strategico: gli individui investono, in modo razionale, le loro energie per creare legami e scelgono le amicizie considerando il costo ed i

benefici di ogni relazione. La struttura della rete è quindi una risultante dell'interazione strategica fra gli agenti.

Esistono due importanti sfide da superare nella costruzione di un modello empirico per la formazione di reti strategiche: la prima è fronteggiare i possibili equilibri multipli che rendono problematica l'identificazione dei parametri strutturali e la seconda riguarda l'intrinseca complessità computazionale legata alla crescita esponenziale del numero possibile di configurazioni della rete in base al numero di giocatori. Quest'ultima problematica rende il calcolo dell'equilibrio per le reti di grandi dimensioni estremamente difficile. Il modello proposto possiede un unico punto di equilibrio e riduce sensibilmente l'onore computazionale. La formazione dei link è sequenziale, ovvero per ogni periodo soltanto un agente è attivo e aggiorna i suoi legami.

Le strategie per decidere la creazione di un legame con un altro agente sono legate agli attributi socio-economici di quest'ultimo. É importante sottolineare come gli incontri fra gli agenti siano frequenti e come ogni agente possa rivedere frequentemente le sue strategie.

Per preservare la trattabilità viene assunto che gli individui non agiscano riflettendo sul fatto che le loro azioni modifichino la forma della rete ma bensì agiscono con una strategia di miglior risposta[4]

[4]Una strategia, nella teoria dei giochi, in cui il giocatore fornisce la risposta che più gli è utile.

Capitolo 2

Reti sociali

Molti sociologi hanno affrontato la tematica della società vista come un grafo in cui i nodi sono rappresentati dagli individui e i gli archi rappresentano le interazioni fra di essi, siano lavorative, sociali o di altra forma.

In questo scenario sono stati approfonditi vari contesti come quello del mercato del lavoro in cui si cerca di capire quale tipo di persona guadagni più soldi o nel processo di assunzione abbia maggiori probabilità di trovare lavoro più facilmente.

Granovetter nel suo studio del 1973 [12] afferma come i legami deboli, di cui si spiega in seguito il significato, assumano un ruolo fondamentale qualora la ricerca di informazioni sia effettuata tramite le interazioni sociali, ed in particolare per il mercato del lavoro. Successivamente Montgomery nel 1991 [18] afferma che in equilibrio gli stipendi non sono determinati dalle abilità dei lavoratori, bensì dai legami posseduti, secondo la regola del

"it's not what you know but who you know"

Come logica conseguenza i gruppi protetti si mostrano soggetti ad appartenere ad una rete sociale che offre meno sbocchi lavorativi degli altri gruppi e quindi a parità di produttività, sono discriminati. Tale affermazione è giustificata dal fatto che le raccomandazioni da parte di un dipendente a favore di una suo conoscente, siano meno costose di un normale processo di assunzione e che inoltre apportano una scrematura aggiuntiva.

Rees [24] Peter Doeringer e Michael Piore [23] affermano come i dipendenti tendano a raccomandare individui dalle caratteristiche simili a loro stessi. Rees dichiara che la reputazione di un dipendente è in gioco quando viene spesa una buona parola per un potenziale candidato ad un posto di lavoro nell'azienda in cui egli lavora.

Nello studio di Montgomery, si definiscono tre tipi di canali attraverso cui viene trovato lavoro:

- **Legami forti** o *strong ties* rappresentano i contatti intra-gruppo ovvero fra individui appartenenti alla stessa "comunità" intesa come una regione in cui gli individui sono strettamente collegati.
- **Legami deboli** o *weak ties* rappresentano i contatti inter-gruppo ovvero fra individui appartenenti a "comunità" differenti. Rappresentano interazioni sociali casuali, non frequenti e che tendono ad attraversare regioni eterogenee, parti distanti della struttura della rete sociale.
- **Canali formali** rappresentano agenzie per l'impiego pubbliche e private, inserzioni su giornali, incontri sindacali di assunzione, e dei servizi di collocamento del college.

In seguito agli studi di Granovetter si è capito quanto sia fondamentale, in una rete sociale, la presenza di legami deboli per la diffusione rapida delle informazioni e che la sostituzione di questi con legami forti, comporta una frammentazione maggiore della società che a sua volta causa un rallentamento nel processo di propagazione. É stato dimostrato inoltre come all'aumentare della proporzione di legami deboli, la disuguaglianza diminuisce e potenzialmente si possa incrementare l'occupazione.

Un esempio concreto di utilizzo della rete sociale per comprendere alcuni fenomeni reali è quello del sociologo Henry Overman che nel 2002 [7], analizza dati sugli adolescenti australiani per scoprire se esistono effetti sul tasso di abbandono scolastico legati alle amicizie vicine.

Overman ha scoperto che la composizione culturale delle proprie amicizie può influenzare il tasso di abbandono e in modo più sorprendente evidenzia come avere delle amicizie con un basso stato socio-economico comporta un più basso tasso di abbandono.

2.1 La segregazione come forma di discriminazione

Van der Leij attraverso un suo studio empirico [28] afferma come non sia tanto la discriminazione quanto piuttosto la segregazione in grado di spiegare il divario salariale fra gli individui.

Nell'analisi dei comportamenti delle minoranze, afferma come i membri appartenenti a queste guadagnino in media meno dei gruppi maggioritari. Per questo si focalizza sulle caratteristiche dei lavoratori appartenenti ai gruppi minoritari.

Egli sottolinea come gli studi economici abbiano da sempre dato poca importanza al contesto sociale ed alla sua influenza sulla realtà dei lavoratori, a differenza della letteratura inerente alla sociologia ed alle scienze politiche.

Per capire se le minoranze siano collegate a trattamenti disparitari, Van der Leij introduce due tipi di misure del contesto sociale: la composizione della comunità locale in gruppi di individui di etnia uguale e la omofilia nella rete sociale.

2.2 Il modello economico di Armengol e Jackson

Questi studiosi assumono che esista un modello economico sottostante semplice a differenza della struttura della rete sociale che può essere complessa. A differenza di Granovetter e Montgomery non fanno distinzioni fra legami forti o deboli ma si basano sulla probabilità di perdere il lavoro e di venire a conoscenza di una offerta di lavoro.

Calvó-Armengol e Jackson [4] osservano come nel mercato del lavoro, se la diffusione delle informazioni avviene mediante una rete sociale, gli individui che posseggono un lavoro con uno stipendio alto siano predisposti ad agevolare la diffusione delle offerte di lavoro. Questo significa che se un individuo è collegato ad altri disoccupati sarà molto probabile che resti senza lavoro a lungo, poiché avrà più difficoltà a ricevere dalla rete informazioni inerenti al lavoro. Tale processo può portare alla diffusione di certi comportamenti come ad esempio l'abbandono della rete sociale a cui si appartiene.

Un individuo collegato ad una persona che abbandona la rete può essere stimolato ad abbandonare anch'egli la rete: venendogli a mancare un canale ricettore di informazioni, percepisce di avere maggiori difficoltà a ricevere offerte di lavoro. Da questo caratteristica che tende a diffondere sulla rete determinati comportamenti partendo da un piccolo sottoinsieme di nodi adepti, si può affermare come all'interno di una rete complessa sia più determinante un intervento mirato ad un gruppo specifico di individui che un'azione generale su nodi scelti a caso. Nel primo caso l'effetto di diffusione può amplificarsi su tutta la rete mentre nel secondo caso spesso si vanifica nel nulla.

g	1 period	2 periods	10 periods	limit
	0.099	0.099	0.099	0.099
	0.176	0.175	0.170	0.099
	0.305	0.300	0.278	0.099

Figura 2.1: Effetti della rete sociale sull'occupazione

Per capire come la rete influenzi la ricerca di un lavoro, i due studiosi definiscono la probabilità che un agente i sia informato di una offerta di lavoro da parte di un agente j nell'equazione 2.1.

Definita a la probabilità di venire a conoscenza di un'offerta un lavoro e definito come un arco $g_{ij} = 1$ se un agente i conosce un agente j (0 altrimenti). Due agenti si conoscono reciprocamente se $g_{ij} = g_{ji}$, la variabile tempo è introdotta grazie al vettore s_t che descrive lo stato di impiego degli agenti al tempo t: se un agente i ha una occupazione alla fine del periodo t allora $s_{it} = 1$ se è disoccupato $s_{it} = 0$.

$$p_{ij}(s)\begin{cases} a & \text{Se } s_i = 0, i = i \\ \frac{a}{\sum_{k:s_k=0} g_{ik}} & \text{Se } s_i = 0, s_j = 0, g_{ij} = g_{ji} = 1 \\ 0 & \text{altrimenti} \end{cases} \tag{2.1}$$

L'equazione 2.1 illustra come la probabilità che un agente i venga a conoscenza di un lavoro da parte di un suo contatto j è inversamente proporzionale al numero di contatti disoccupati che possiede.

In figura 2.1 si può notare come a seconda di quale rete sia trattata aumenti o resti costante la probabilità di trovare lavoro con il variare del tempo. Tutti i nodi raffigurati rappresentano agenti con un impiego fisso. Nella prima riga si osserva come in una rete totalmente sconnessa la probabilità di venire a conoscenza di un'offerta di lavoro nel tempo rimanga invariata perché nessun nodo ha legami da cui poter ricevere informazioni.

Nell'ultima riga della figura viene mostrata una rete connessa con legami fitti, un nodo i nel tempo aumenta le possibilità di impiego poiché avendo un insieme di contatti γ incrementa la possibilità di ricevere offerte di lavoro dagli elementi di γ che sono riusciti a trovare lavoro prima di lui.

I dati riscontrati sopra sono la logica conseguenza dell'equazione 2.1 che afferma come la probabilità di un agente i di venire a conoscenza di un'offerta di lavoro j aumenti proporzionalmente al numero di contatti che possiedono già un lavoro e che quindi sono predisposti a diffondere l'informazione.

Gli studi di Calvo-Armengol e Jackson si relazionano con la teoria della discriminazione, vista in Shelling, in cui grazie alle intrinseche preferenze individuali le persone si riuniscono in gruppi omogenei, causando quel fenomeno chiamato segregazione e che spiega le disuguaglianze nel trattamento a parità di condizioni. Essi offrono una visione alternativa per spiegare la mancanza di parità di trattamento: attribuiscono l'influenza del successo economico degli individui alla rete sociale e sostengono che la disuguaglianza sia il risultato di differenti storie individuali legate alla rete di conoscenze.

2.3 Omofilia

Con *omofilia* si intende quel processo che tende a formare all'interno di una rete sociale, gruppi di individui omogenei sotto il punto di vista demografico (età, sesso, razza) e culturale (attitudini, aspirazioni, educazione).

Platone nel 380 a.C., nella sua opera chiamata Fedro, enuncia un proverbio di antiche origini:

"il coetaneo si diletta del coetaneo" (240a)

La consuetudine di trovare piacere a stare in contatto con persone affini in età ma anche culturalmente non possiede età.

I primi studi sistematici furoni fatti in sociologia in contesti sociali di piccoli gruppi (bambini nelle scuole, universitari, vicini di casa).

Un indice di omofilia può essere matematicamente espresso come la frequenza relativa con cui gli individui interagiscono socialmente con le altre persone simili a loro.

Data una popolazione di dimensione N divisa in due o più gruppi sulla base delle caratteristiche personali ed in cui N_c indica la dimensione di un gruppo di individui di tipo c, w_c la proporzione di individui di tipo c nella popolazione globale (N_c/N), e sia s_i il numero di legami formati da individui i con persone dello stesso tipo e d_i il numero di legami formati con individui di tipo differente, allora un indice di omofilia può essere definito come

$$h_i = \frac{s_i}{s_i + d_i}$$

In seguito allo studio delle reti complesse sono state introdotte misure come l'*assortative mixing* per misurare nella rete i nodi con caratteristiche simili [20]. In presenza di questo fenomeno i nodi della rete tendono a possedere archi solo con nodi dello stesso tipo, mentre il comportamento inverso, ben più raro, è chiamato *disassortative mixing*.

Esistono due tipi di approcci per calcolare l'assortative mixing in base al modo di classificare i nodi: quello *discreto* in base alle loro caratteristiche intrinseche (ad esempio al tipo) e quello *scalare* in base a delle proprietà scalari. Un'altra misura quantitativa che esprime il grado di omofilia in una rete viene chiamata *coefficiente di assortativity* R È un numero reale compreso fra -1 e 1 ed a seconda del valore che assume esprime la natura della rete

$$R = \begin{cases} 0 & \text{Se la rete non possiede assortative mix} \\ +1 & \text{Se la rete possiede un assortative mix perfetto} \\ < 0 & \text{Se esiste disassortative mix} \\ -1 & \text{Se esiste disassortative mix perfetto} \end{cases}$$

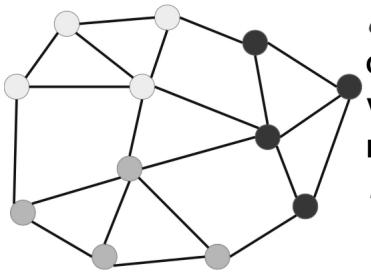
Discreto

Come detto in precedenza, viene considerata una rete in cui esistono nodi di tipo differente e con tipo si intendono caratteristiche intrinseche del nodo. Nell'approccio discreto viene calcolato un valore e_{ij} che esprime la frazione di archi della rete che connettono nodi del *tipo* i con nodi del *tipo* j.

In seguito vengono calcolate le variabili $a_i = \sum_j e_{ij}$ che rappresenta la frazione degli archi uscenti da nodi di tipo i e $b_j = \sum_i e_{ij}$ che rappresenta la frazione degli archi entranti di tipo j.

R è calcolato in questo approccio come

$$R = \frac{\sum_i e_{ii} - \sum_i a_i b_j}{1 - \sum_i a_i b_i}$$



e_{ij}	G	V	R	a
G	5/21	2/21	2/21	3/7
V	2/21	5/21	2/21	3/7
R	2/21	2/21	5/21	3/7
b	3/7	3/7	3/7	

Figura 2.2: Assortative Mix Discreto su grafo

Al numeratore viene rappresentata la differenza tra la frazione degli archi che collegano nodi dello stesso tipo e la frazione degli archi che collegano nodi di tipo differente. Al denominatore è rappresentata la frazione di tutti gli archi che non collegano nodi dello stesso tipo. Se il numeratore diventa negativo vuol dire che gli archi che collegano nodi di tipo differente sono maggiori degli archi che collegano nodi di tipo simile, viceversa se il numeratore è positivo vuol dire che la rete possiede in maggioranza soltanto archi che collegano nodi dello stesso tipo.

Nell' esempio in figura 2.2, il coefficiente di assortativity è R = 0.3636

Scalare

Come detto in precedenza, viene considerata una rete in cui i nodi possiedono proprietà scalari, ovvero ad ogni nodo è associato un valore. Nell'approccio scalare vengono calcolate le variabili e_{xy} che esprime la frazione degli archi della rete che connettono nodi con valore x a quelli con valore y, $a_x = \sum_y e_{xy}$ che rappresenta la frazione degli archi uscenti da nodi con valore x e $b_y = \sum_x e_{xy}$ che rappresenta la frazione degli archi entranti in nodi con valore y. Il coefficiente di assortativity R è calcolato come

$$R = \frac{\sum_{xy} xy \cdot (e_{xy} - a_x b_y)}{\sigma_a \sigma_b}$$

Il significato dell'indice è lo stesso visto nel modo discreto.

2.4 Le *esternalità* della rete e le dinamiche delle disuguaglianze di gruppo

Lo studioso Kim Young Chul nel 2009 [14] cerca di capire le cause che comportano il successo economico di una rete sociale, affrontando anche la questione psicologica.

Osservando il sistema lavorativo come un susseguirsi di periodi quali ad esempio l'acquisizione di abilità, il tutoraggio, la ricerca di un lavoro ed i relativi contatti in seguito ed infine la canalizzazione delle informazioni, Kim

definisce le *esternalità* di una rete tutti quei processi al di là del periodo educativo ed afferma che siano fondamentali per la creazione di disuguaglianze all'interno dei gruppi.

Ad esempio afferma come la visione ottimista del futuro di un gruppo protetto possa influenzare gli investimenti in formazione poiché vi è la speranza di un miglioramento della propria posizione sociale. Inoltre durante il periodo di formazione (teoria dello studioso Loury, 1977), un individuo viene a contatto con dei vincoli sulla sua formazione, sulla nutrizione, sull'assistenza medica, sul dopo scuola dei figli, sui modelli preconfigurati dei ruoli che influenzano la sua visione della vita.

Kim individua due tipi di esternalità: quelle che operano durante il periodo educativo e quelle che agiscono nella vita di un lavoratore.

Le prime tendono ad agire nel *passato* ovvero per cambiare lo stato di un gruppo è necessario aver investito in formazione nel passato.

Le seconde tendono ad agire nel *futuro* ovvero se voglio influenzare gli investimenti in formazione di un gruppo devo modificare i benefici futuri accumulati dall'acquisizione di abilità. Questa seconda caratteristica offre agli individui la possibilità di agire insieme per migliorare o deteriorare la qualità di una rete sociale. Ad esempio se un gruppo minoritario con una situazione difficile produce una generazione di individui che crede in un futuro migliore e si impegna nell'acquisire abilità, questo processo può portare, se le generazioni successive continuano su questa visione ottimistica, a migliorare la loro situazione; parimenti può essere considerata la situazione inversa in cui un gruppo in una ottima posizione a causa di visioni pessimistiche può peggiorare la sua situazione (eccezione fatta per in cui casi la situazione è messa troppo bene o troppo male).

Una interessante caratteristica dell'economia multi gruppo è definita come *trappola sociale* ovvero quando un gruppo mantiene un tasso di alta formazione ed un altro è "intrappolato" dalle scelte passate in cui è stata favorita la scarsa qualità della formazione e che comporta l'impossibilità da parte del gruppo minoritario di uscire dalla situazione difficoltosa senza l'intervento governativo. Allo stesso tempo afferma che in un primo stato di sviluppo di una economia, la situazione di ingiustizia può fare decollare la crescita economica, perchè almeno i gruppi in condizione di povertà sono stimolati a sviluppare le loro abilità.

Kim afferma come sia possibile uscire dalla tradizionale visione del mercato imperfetto per definizione ed introdurre l'idea di uguaglianza che, secondo lui, ha effetti positivi sullo sviluppo di una economia senza vincoli sul credito infondati o in una società con un servizio scolastico pubblico.

L'equilibrio si trova redistribuendo equamente il capitale sociale[1], inteso come la media del capitale umano[2] di una rete sociale: questo può inco-

[1] indica l'insieme delle relazioni interpersonali formali ed informali essenziali anche per il funzionamento di società complesse ed altamente organizzate

[2] è l'insieme di conoscenze, competenze, abilità, emozioni, acquisite durante la vita da

raggiare i gruppi sociali ad incrementare i loro tassi di investimento nella formazione e farli uscire dalla trappola sociale in cui sono caduti.

Lo stato può agire attraverso due tipi di *azioni positive*

- sussidi di formazione,
- sistema delle quote.

Nel primo caso il governo può imporre delle tasse ai gruppi avvantaggiati e trasferire risorse a quelli svantaggiati con lo scopo di aumentare l'investimento in formazione (vengono infatti abbassati i costi per l'educazione).

Nel secondo caso invece vengono fissate delle quote per le posizioni di lavoro altamente qualificate che vengono concesse ai gruppi minoritari con lo scopo di ridurre la disparità fra i gruppi.

Kim individua (citando Goldin e Katz, Abramovitz e David) nel capitale umano *" la chiave del motore nella crescita dell'economia moderna"*, affermando che l'accumulo di questo capitale intangibile sta rimpiazzando l'importanza dall'accumulo di capitale economico nella fase iniziale di una rivoluzione industriale.

un individuo e finalizzate al raggiungimento di obiettivi sociali ed economici, singoli o collettivi.

Capitolo 3

Processo di data mining

Il processo di data mining si compone di diversi passi concettuali che ha portato alla concepimento del sistema:

1. l'analisi dei requisiti,
2. la definizione di segregazione,
3. la definizione di gruppi protetti,
4. la descrizione della rete sociale su cui si agisce,
5. la scelta dell'algoritmo per individuare le comunità,
6. la scelta degli indici utilizzati,
7. l'architettura del sistema e la descrizione delle sue componenti.

3.1 Motivazione del progetto

La Comunità Europea si sta muovendo per cercare di vietare tutte le forme di discriminazione secondo alcuni criteri, questa volontà non ha ancora trovato utili strumenti per individuare la presenza di tale fenomeno all'interno di grandi quantità di dati strutturati.

Come esposto in precedenza esiste un legame forte fra segregazione e discriminazione. Una comunità viene discriminata quando il comportamento adottato dalla società nei suoi confronti è diversificato rispetto alle altre in base a determinati attributi (età, sesso, posizione sociale, colore e cosi via), avviene una disparità di trattamento a sfavore delle categorie deboli.

Un individuo segregato ovvero separato dalla massa, non può usufruire di tutte le possibilità che il suo status gli consentirebbe.

Come si è visto in letteratura, la segregazione in passato è stata calcolata all'interno di unità organizzative fisiche (scuole, città, quartieri), ben definite ed in piccole realtà in cui erano coinvolti pochi individui e conseguentemente

relazioni, attraverso il lavoro *manuale* [1] di sociologi ed economisti fino a poco tempo fa, causa la mancanza di strumenti automatizzati.

Per attualizzare questo tipo di analisi in cui i nodi e gli archi diventano decine o centinaia di migliaia, si presenta il problema della *scalabilità* che limita l'utilizzo delle tecniche classiche. Solamente quindi attraverso specifici modelli implementati con algoritmi ad hoc si può ovviare a tale tipo di impedimento, sfruttando la velocità di calcolo degli elaboratori e la memoria di cui dispongono.

Per questo motivo la tesi verte sull'ideazione e creazione di un modello sperimentale di calcolo della segregazione per una specifica rete sociale economica reale, fornendo uno strumento di analisi del comportamento di grandi quantità di dati strutturati. Attraverso tale tipo di strumento si offre la possibilità di individuare forme accentuate e latenti di segregazione.

3.2 Analisi dei requisiti

Lo scopo di questa sezione è di illustrare le funzionalità che il sistema deve avere:

- un metodo ETL^2 per estrarre dati da una sorgente file, trasformarli e caricarli nel sistema,
- creazione della rete sociale,
- scrittura su file della rete per analizzarla,
- calcolo delle comunità,
- calcolo di una misura di segregazione per le unità trovate,
- scrittura su file dei risultati.

Oltre a queste, ci siamo posti la possibilità di fornire il calcolo di più misure di segregazione.

Nella sezione successiva è stato specificato che cosa si intende per segregazione per definire cosa le misure devono rappresentare.

3.3 Definizione della segregazione

Come definizione di segregazione si è scelto quella del dizionario dell'istituto Treccani (versione coincisa)

"*Isolare un individuo, o un gruppo di individui, dalla comunità di cui fa parte, tenendolo lontano da questa.*"

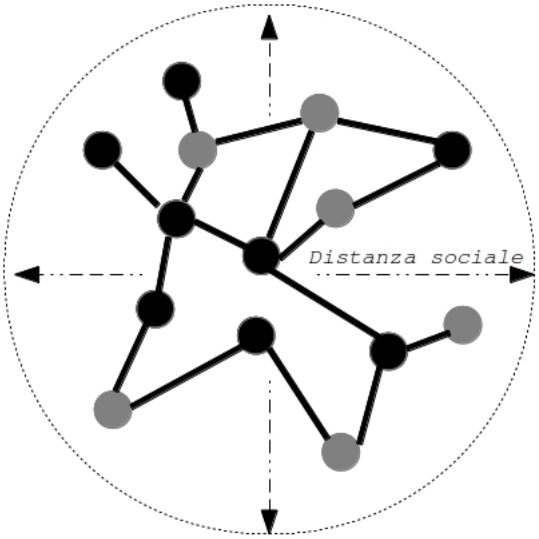


Figura 3.1: Rappresentazione di una comunità

Come si può osservare dalla figura 3.1, intuitivamente una comunità è definita come un insieme di nodi "vicini", i quali rispettano una distanza sociale (definita a discrezione).

Le reti sociali reali, figura 3.2, sono composte tipicamente da sotto-comunità locali, nel nostro caso vengono chiamate *unità organizzative*.

In rispetto a quanto appena detto, sorge la necessità di capire il tipo di rete da analizzare per poter definire la distanza sociale che distingue quando un nodo appartiene ad un'unità organizzativa o ad un'altra.

Esistono vari criteri di similarità che possono essere utilizzati, da quelli basati sulla distanza (quantificata in una qualche unità di misura, come ad esempio quella metrica) a quelli basati sulla similarità degli individui (simile per razza, religione, stipendio e via dicendo) e molti altri.

La segregazione è stata calcolata

all'interno di ogni unità organizzativa in funzione di un gruppo protetto.

La scelta di come verranno individuate le *unità organizzative* all'interno della rete è definita e giustificata nella sezione 3.6.

Ogni sottocomponente all'interno della rete individua una micro realtà sulla quale andare a calcolare le misure spiegate nella sezione 3.8.

[1] tramite interviste faccia a faccia ed altre tecniche non informatizzate
[2] extract, transfert, load

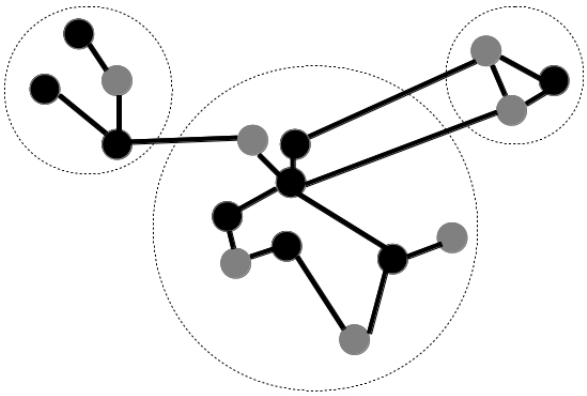


Figura 3.2: Struttura delle comunità in una rete complessa

La semantica di un' unità organizzativa dipende in modo imprescindibile dal tipo di contesto:

- in una popolazione di studenti, possono essere individuate andando a scegliere come metrica la distanza spaziale (l'indirizzo del luogo di studio o la classe di appartenenza) poiché le comunità scolastiche in cui si vuole osservare la segregazione sono le scuole o le classi.

- in una popolazione di lavoratori, possono essere viste come il sottoinsieme di aziende che intrattengono rapporti di lavoro. I criteri di selezione dei rapporti possono variare dall'intensità dei rapporti (numero di dipendenti in comune fra due aziende) alla profondità del contatto (direttamente collegate oppure il livello di contatto: lavoro per conto terzi). É ragionevole pensare che se due aziende intrattengono rapporti di lavoro o possiedono dipendenti comuni, possano considerarsi parte della stessa comunità. Un insieme di persone che si conoscono o vengono a contatto durante l'orario di lavoro possono essere considerate appartenenti della stessa unità organizzativa.

- in un contesto di concessione di mutui per individuare una comunità possono essere scelti criteri misti come residenza (e quindi distanza spaziale) ed interazioni sociali in comune, poiché una comunità *è un insieme di individui che condividono lo stesso ambiente fisico e tecnologico, formando un gruppo riconoscibile, unito da vincoli organizzativi, linguistici, religiosi, economici e da interessi comuni.* (cit. Wikipedia)

Appare evidente la forte influenza esercitata dal ricercatore attraverso il suo modo di interpretare e definire le unità organizzative. Le sue decisioni possono influenzare il risultato, poiché il modo arbitrario di scegliere le comunità porta ad influenzare ogni tipo di misura utilizzata. Tale tipo di procedimento risulta inevitabile quando la realtà dei dati impossibilita

un approccio oggettivo del ricercatore. Spesso lo scenario analizzato mette lo studioso di fronte alla necessità di definire parametri con il suo metro di giudizio ed appare naturale che la qualità dei risultati sia influenzata dall'esperienza accumulata nel passato.

3.4 Definizione di gruppo protetto

La segregazione è misurata in funzione di un gruppo protetto ma prima di definire cosa si intende con gruppo protetto è necessario scegliere gli attributi sensibili da utilizzare:

- sesso (attributo nominale),
- età (attributo numerale).

Questa scelta è stata fatta poiché le informazioni sugli amministratori presentano un alto tasso di valori mancanti per i residui attributi ad esempio il codice di avviamento postale della residenza. I due attributi usati hanno permesso di minimizzare gli individui scartati per mancanza di valori.

Un gruppo protetto GP_{xy} è stato definito come

"*un insieme di amministratori con un vincolo sull'età x e uno sul sesso y.*"

É stato individuato un gruppo protetto per ogni valore dell'età e per ogni valore del sesso in modo da poter esprimere una ampia visione della realtà. Di seguito si riportano i vincoli sugli attributi.

3.4.1 Vincoli sull'età

Sono stati utilizzati due tipi di approcci

- *taglio ottimo binario*
  - *Modo 1*
  - *Modo 2*
- *media mobile - Modo 3*

Di seguito si illustrano nel dettaglio, tralasciando per il momento l'attributo sesso.

Approccio a taglio ottimo binario - Modo 1

GP_x viene identificato come quell'insieme di amministratori per cui ogni individuo i possiede una età maggiore o uguale a x

$$GP_x = \{\text{amministratore } i | anni_i \geq x\}$$

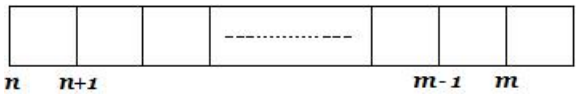


Figura 3.3: Indice di isolamento sull'età

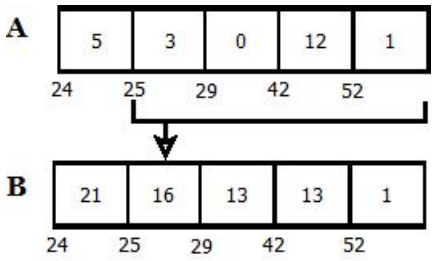


Figura 3.4: Calcolo della dimensione dei gruppi per l'età - Modo 1

Nella figura 3.3 la cella m rappresenta il valore dell'indice per il gruppo protetto GP_m i cui individui hanno età $\geq m$, mentre la cella $m-1$ contiene il valore dell'indice calcolato per il gruppo protetto i cui individui hanno età $\geq$ *m-1*. La differenza di valore fra le due celle rappresenta il contributo dell'età m al valore dell'indice confrontato con il resto della comunità.

L'esempio nella figura 3.4 mostra due strutture A e B che contengono una cella per ogni età rappresentata dagli individui del dataset (in questo caso 24, 25, 29, 42, 52 anni).

Nella struttura A per ogni cella indicizzata i viene memorizzata la frequenza assoluta degli individui con l'età uguale a i: la prima cella di A contiene tutti gli individui che hanno un'età di 24 anni.

La struttura B per ogni cella indicizzata i contiene il gruppo protetto GP_i, la prima cella di B rappresenta tutti gli individui che hanno una età ≥ 24 anni. La seconda cella della struttura A indica che esistono tre individui di 25 anni mentre la struttura B nella solita cella indica che gli individui di età ≥ 25 sono 16.

Questo approccio è stato scelto per analizzare i gruppi protetti di amministratori anziani. Partendo dall'insieme di amministratori con l'età più grande e di volta in volta aggiungendo membri più giovani si è cercato di capire come l'insieme delle persone più "datate" influenzasse i valori degli indici studiati.

Approccio a taglio ottimo binario - Modo 2

Questo tipo di approccio è stato scelto per esprimere l'indice per i gruppi protetti di amministratori giovani.

$$GP_x = \{\text{amministratore } i | anni_i \leq x\}$$

Approccio media mobile - Modo 3

Questo tipo di approccio utilizza una tecnica di *smoothing*, con lo scopo di evitare i tipici fenomeni legati alle entità *outlier*. Gli amministratori o troppo giovani o troppo anziani possono alterare i valori delle misure utilizzate, con la conseguenza di errate interpretazioni distanti dalla realtà.

Il concetto alla base dello smoothing è di fondere dei valori vicini fra loro, nel nostro caso di includere nello stesso gruppo gli amministratori con età simile

$$GP_x = \{\text{amministratore i} \mid x - WS \leq anni_i \leq x + WS\}$$

WS è definito come una costante che indica la dimensione della finestra entro cui considerare un amministratore simile d'età ad un altro.

Fissando ad esempio WS=3, GP_{38} individua gli amministratori con una età compresa nell'intervallo [35;41].

Individuare un gruppo protetto identificato da un anno di nascita (ad esempio il gruppo protetto di quelli nati nell'anno solare 1975) non è un approccio preso in considerazione perchè si pensa possa essere affetto da overfitting [3].

3.4.2 Vincoli sul sesso

Tralasciando l'età, un gruppo protetto è stato individuato come

$$GP_y = \{\text{amministratore } i | sesso = y\}$$

Per cui un gruppo protetto GP_{maschi} rappresenta l'insieme di amministratori di sesso maschile.

3.4.3 Numero dei gruppi protetti di analisi

Ricordando la definizione di gruppo protetto GP_{xy} definito come

"*un insieme di amministratori con un vincolo sull'età x e uno sul sesso y.*"

Si calcolano le misure di segregazione analizzando tutti i gruppi protetti definiti per ogni valore dell'età e per ogni valore del sesso.

Nella sezione 3.7 si è definita l'età massima $m = max(x) = 65$ e l'età minima $n = min(x) = 15$, i possibili valori all'interno di questo intervallo sono 50 (in generale $m - n$).

Per quanto riguarda il sesso siamo interessati ai gruppi protetti di femmine, maschi e misti, ovvero senza considerare il sesso. Inoltre bisogna aggiungere che esistono 3 modi di concepire un gruppo, come si è visto in precedenza.

[3] sovra-adattamento ad un modello e che quindi porterebbe a risultati non confrontabili fra reti sociali simili.

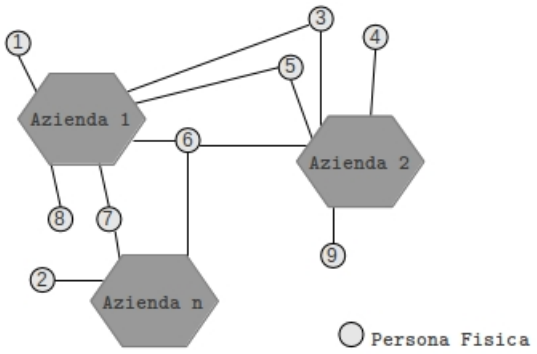


Figura 3.5: Rete reale analizzata

Combinando i vincoli sul sesso, sull'età e sugli approcci utilizzati, il numero totale di gruppi protetti studiati è

$$(m - n) \cdot 3 \cdot 3 = 450$$

3.5 Definizione della rete sociale di analisi

Uno dei requisiti del sistema è di calcolare la rete sociale di analisi, ecco quindi che risulta necessaria la sua definizione.

É stata presa in considerazione una rete sociale economica che rappresenta il mondo del lavoro di migliaia di amministratori di aziende italiane di cui si forniranno i dettagli nella sezione 4.2.1.

In un primo momento, figura 3.5, i dati sono rappresentati come un insieme di aziende scorrelate che posseggono un numero variabile di amministratori. I nodi a forma esagonale rappresentano le aziende mentre i nodi a forma circolare sono gli amministratori. Esiste una relazione fra un esagono A ed un cerchio B se l'individuo B amministra oppure ha amministrato in passato l'azienda A.

In un secondo momento per identificare le unità organizzative è stato necessario costruire una rete omogenea di nodi. In cui i nodi rappresentano le aziende e gli archi sono creati se esistono amministratori in comune.

Il concetto è schematizzato nella figura 3.6: la parte sinistra rappresenta i dati di partenza, definiti come una rete eterogenea in cui sono presenti sia gli amministratori (i puntini neri) che le aziende; la parte destra rappresenta la rete così come sarà trattata dal software, un insieme omogeneo di nodi che rappresentano le aziende trattate.

Durante la fase di trasformazione da rete eterogenea a rete omogenea, una unità organizzativa è stata definita come

"*un insieme di aziende che possiedono amministratori in comune.*"

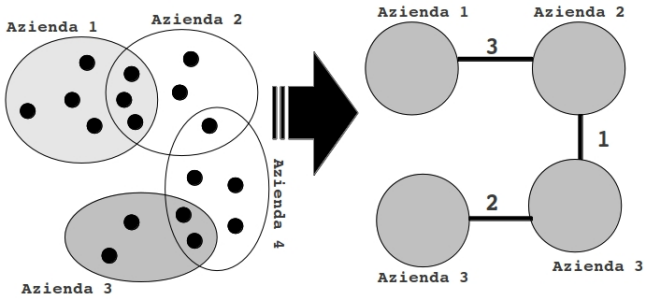


Figura 3.6: Trasformazione da rete eterogenea ad omogenea

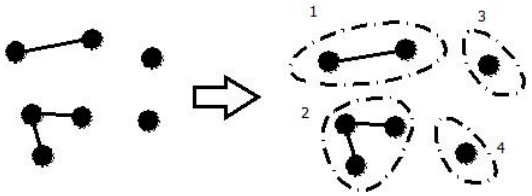


Figura 3.7: Algoritmo di individuazione delle comunità

3.6 Calcolo delle Comunità

Dato che un requisito del sistema è l'individuazione delle unità organizzative all'interno della rete sociale, di seguito si riporta l'algoritmo utilizzato ed il motivo per cui è stato scelto.

In primo luogo bisogna precisare che il passo concettuale precedente la scelta dell'algoritmo è stata la creazione della rete sociale e l'analisi delle sue caratteristiche. É importante sottolineare questo passo, poiché è solo attraverso una profonda conoscenza del dominio di analisi che si possono giustificare le scelte fatte.

É stato utilizzato un unico algoritmo che individua le componenti debolmente connesse della rete. Una componente debolmente connessa è definita come il sottoinsieme massimale di vertici nel quale tutte le coppie di vertici sono raggiungibili l'una dall'altra nel sottostante sottografo non orientato.

Questo algoritmo è stato scelto poiché la rete sociale risulta formata da componenti debolmente connesse di piccole dimensioni, per i dettagli la sezione 3.7. Questo tipo di caratteristica non consente di applicare ulteriori criteri di suddivisione poiché andrebbe perso il significato di comunità.

Come si vede nella figura 3.7, la parte sinistra della figura rappresenta la rete e gli archi che la compongono, nella parte destra si vedono le quattro componenti individuate dall'algoritmo.

L'algoritmo utilizzato impiega un tempo di $\mathcal{O}(|V|+|E)$ dove $|V|$ rappresenta il numero di vertici ed $|E|$ il numero di archi.

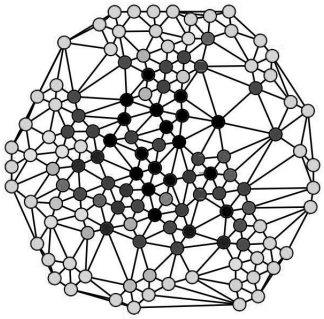


Figura 3.8: Betweenness centrality

3.6.1 Misure teoriche

Nella letteratura scientifica inerente all'analisi delle reti sociali si discutono varie misure di centralità di un vertice all'interno di un grafo che determina la relativa importanza dello stesso all'interno della rete a cui appartiene. Alcune di queste possono essere utilizzate per individuare comunità attraverso alcune tecniche che tendono a separare i nodi distanti fra loro.

Di seguito ne vengono illustrate alcune a carattere illustrativo:

- "*betweenness centrality*", una funzione $g(v)$ che per ogni nodo v appartenente ad una rete rappresenta la sua centralità nella rete: un misura che conta il numero di cammini minimi che attraversano tale nodo, come si vede dall'equazione 3.1, in cui il denominatore della sommatoria è il numero di cammini minimi dal nodo s al nodo t, mentre il numeratore è il numero di quei cammini minimi che passa attraverso il nodo v (nella figura 3.8 i nodi scuri rappresentano i nodi con più alti valori di betweenness e viceversa i nodi bianchi).

$$g(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st(v)}}{\sigma_{st}} \tag{3.1}$$

- "*closeness centrality*", una una funzione $f(v)$ che esprime quanto un nodo v sia distante da tutti i suoi vicini, considerata come l'inverso della somma delle distanze da tutti gli altri nodi. Più v è centrale, minore di conseguenza dovrà essere $f(v)$, in qualche maniera può esprimere quanto veloce una informazione ci metta ad essere diffusa sull'intera rete, partendo dal v.

- "*eigenvector centrality*", è una misura dell'influenza di un nodo sulla rete, in questo modo sono assegnati dei punteggio per ogni nodo v legati al fatto che le connessioni con i nodi di alto punteggio contribuiscono maggiormente al punteggio di v stesso.

Nonostante esistano varie misure che possono in linea teorica aiutare a suddividere la rete in sottocomponenti con proprietà simili, nell'applicazione al caso di studio reale preso in esame, applicare questo tipo di misure a livello concettuale non assume significato.

Il motivo come accennato in precedenza risiede nella concezione di comunità: unità organizzative con un numero massimo di nodi pari al 0.05% del totale, a nostro parere, non devono subire ulteriori processi di suddivisione.

La letteratura è utile quando viene recepita come formazione e non come informazione acquisita acriticamente, tale da essere applicata rigidamente come un insieme di assiomi.

Per verificare che la logica utilizzata possa avere un fondamento di correttezza, nel paragrafo successivo viene descritta una tecnica che si ispira alle misure teoriche della letteratura e che sarà applicata durante la fase di creazione della rete.

"*Learning is experience. Everything else is just information.*"
(Albert Eistein)

3.6.2 Misura applicata

Come si vede nella figura 3.6, le aziende hanno legami con altre aziende con un peso degli archi che indica quanto i legami fra le due imprese siano forti. Avendo analizzato la struttura della rete reale oggetto di analisi, descritta nella sezione 3.7, si è osservato come tale network sia caratterizzata da un alto numero di piccole componenti debolmente connesse che apparentemente non offre spunto per ulteriori suddivisioni in comunità.

Non esistendo alcun tipo di *componente gigante* risulta inappropriato implementare metodi che sconnettano ulteriormente la rete. Nonostante il ragionamento appena fatto, per capire se la forza dei legami incida nel calcolo delle unità organizzative in modo tale da ripercuotersi sui valori delle misure di segregazione studiate, è stato studiato un semplice meccanismo di controllo sugli archi.

Definito φ come

"*il numero di amministratori che due aziende hanno in comune*"

tale valore indica la forza del legame che unisce due aziende.

Durante la fase di costruzione della rete, si veda l'algoritmo 2 nella sezione 4.3, un semplice controllo sulla forza del legame fra due aziende tramite φ, ha permesso di creare una rete in cui il peso minimo degli archi fosse $\geq \varphi$.

Tale metodo permette di eliminare dalla rete quei *legami deboli* che tanto nella analisi delle reti sociali sono stati oggetto di studio. In un certo senso è come se si eliminassero gli archi con maggiore "*betweenness centrality*" poiché rappresentano nella rete reale collegamenti deboli e che uniscono realtà distanti fra loro, nella letteratura definiti come *ponti locali*.

φ	Archi	% del totale	Componenti connesse
1	2190	60,43	66.140
2	692	19,09	67.472
3	247	6,82	67.765
4	212	5,85	67.854
≥ 5	283	7,81	67.864

Tabella 3.1: Archi nelle rete di aziende informatiche

3.7 Descrizione delle reti sociali di analisi

I dati manipolati, discussi nella sezione 4.2.1, hanno portato alla creazione di una rete sociale di lavoro, ovvero un grafo in cui i nodi sono le aziende e gli archi vengono definiti se due aziende hanno almeno φ dirigenti in comune.

Sono state studiate le reti delle aziende nel settore informatico e nel settore finanziario, per capire se il contesto lavorativo fosse una determinante di comportamenti differenti, di seguito vengono descritte le loro caratteristiche.

3.7.1 Rete di aziende di Informatica

La rete possiede 67.989 nodi e 3.624 archi, il 94% dei nodi hanno un grado uguale allo zero (con $\varphi \geq 1$), mentre solamente il 6% dei nodi possiede legami con altre aziende, ovvero ha una governance comune.

Il massimo grado è 35 ed è posseduto da un solo nodo, esiste un nucleo di aziende estremamente piccolo in confronto al numero totale di aziende ma che è strettamente collegato, infatti la comunità più grande è formata da circa 30 nodi con grado vicino a 30, corrisponde allo 0,05% della rete.

Il numero di componenti connesse è esposto nella tabella 3.1 in relazione al valore φ ovvero il numero di amministratori in comune.

Il diametro della rete vista come la massima eccentricità dei vertici del grafo è di 20, la lunghezza media di un cammino è 4 mentre il coefficiente medio di clustering è 0.72.

La rete studiata mostra chiaramente uno scenario in cui la gran parte delle aziende nel contesto informatico sono scollegate dal punto di vista amministrativo, nella tabella si vede come già nella prima riga in cui i legami fra aziende sono deboli (solo un amministratore in comune) le componenti sconnesse siano circa il 97.5 % dei nodi, e che le comunità aumentino in maniera relativamente bassa all'aumentare di φ.

A nostro parere questa caratteristica descrive uno scorcio del panorama aziendale fatto di microimpresa, spesso individuali o a carattere familiare e che può riscontrarsi a pieno nel contesto italiano.

Nella fotografia della rete, in figura 3.9, si mostrano i nodi le cui dimensioni sono in proporzione al numero di amministratori che possiede (più

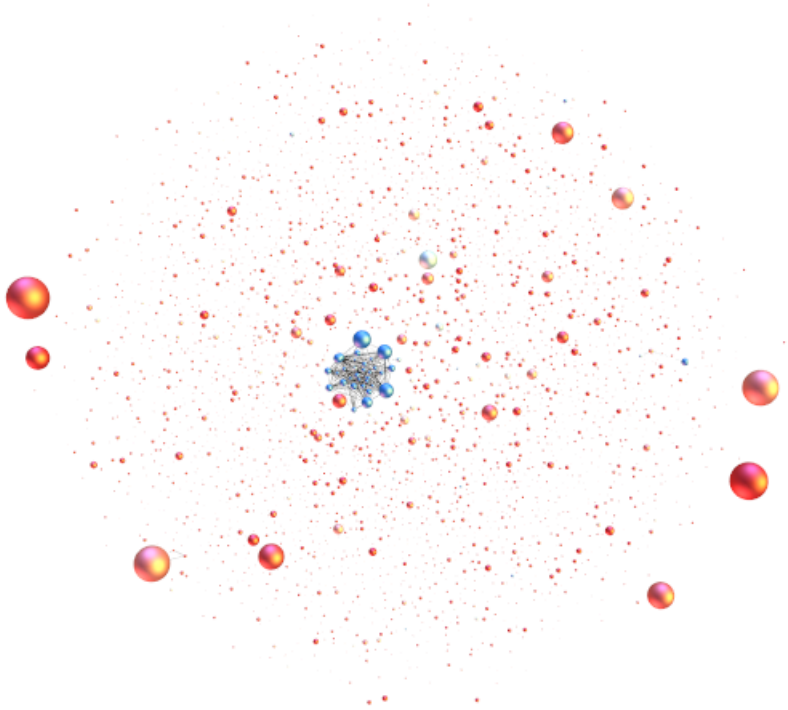


Figura 3.9: Rete delle aziende informatiche

grandi sono e più ne possiedono), ed il loro colore indica il grado del nodo (le gradazioni rosse indicano nodi con basso grado, mentre il blu identifica i nodi con alto grado). L'immagine è una rappresentazione della rete filtrando le associazioni fra aziende che hanno in comune almeno 3 dirigenti. Le aziende con più alto grado possiedono fra loro molti legami, formando un'unica componente fortemente connessa.

Nonostante la rete non sia complessa, mostra alcune somoglianze, ad esempio la distribuzione del grado dei nodi è a "coda lunga" ed evidenzia come pochi nodi abbiano un grado alto (quindi siano collegati a molte aziende) e molti nodi abbiano un grado prossimo allo zero. La densità della rete è zero ed ipotizza una comunicazione intra-gruppi inesistente, in realtà risulta essere solo una supposizione non dimostrabile, poiché sarebbe necessario studiare la frequenza dei contatti fra i nodi e ciò non possibile poiché non si dispone dei dati necessari.

Descrizione della popolazione

Gli amministratori analizzati in questo scenario sono stati 100.456 di cui il 71% sono maschi, bisogna anche annotare che sono stati scartati per mancanza di dati 13.168 individui che rappresentano l'11,5% della popolazione complessiva, una quota elevata a causa di inconsistenze nei dati di origine, valori mancanti e per un filtraggio sull'età tale per cui abbiamo considerato solo gli individui compresi nell'intervallo [15;65].

Il motivo per cui si è scelto tale intervallo di età risiede nel campionamento statistico tipico dell'Istat. L'Istituto nazionale di statistica è solito analizzare il mondo del lavoro considerando quel intervallo di età.

Nella figura 3.10 si osserva come la distribuzione degli individui per età vada ad accostarsi ad una gaussiana.

Si evidenzia come i maschi siano nettamente superiori alle femmine.

3.7.2 Rete di aziende di Finanza

La rete possiede 10.965 nodi e 7.964 archi, il 70% dei nodi hanno un grado uguale allo zero (con $\varphi \geq 1$) ed è una misura che indica come la rete delle aziende nel contesto della finanza siano simili a al contesto dell'informatica. Nella figura 3.11 appare la rete privata dei nodi con grado zero, la dimensione di un nodo è proporzionale al numero di amministratori che possiede, mentre l'intensità del colore è proporzionale al grado, più intenso corrisponde a un grado più alto. Il diametro della rete è 19 e la densità è zero.

La distribuzione del grado dei nodi è a "coda lunga" proprio come la rete di aziende informatiche. La tabella 3.2 evidenzia la differenza con la rete precedente poiché i legami fra le aziende sono decisamente maggiori nonostante il numero di nodi sia più di 6 volte inferiore a quello della rete informatica (7 archi ogni 10 nodi contro i 5 archi ogni 100 nodi della rete di aziende informatiche). Il grado medio dei nodi è 0.81.

Si nota come esista un forte incremento delle componenti debolmente connesse all'aumentare di φ e questo indica una forte commistione fra le varie aziende. Questo tipo di caratteristica è riconosciuta nel contesto finanziario dove sono maggiori le possibilità di trovare un amministratore che risieda in più consigli di amministrazione.

Questo tipo di rete sembra essere un motivo valido per cui è stato scelto di implementare la tecnica per il calcolo delle comunità discussa nella sezione 3.6.2. La lunghezza media di un cammino è di 6.5 e ciò è un dato che ricorda lo scenario dello "*small world*" ovvero indica la possibilità di raggiungere parti della rete distanti in media visitando 6 nodi.

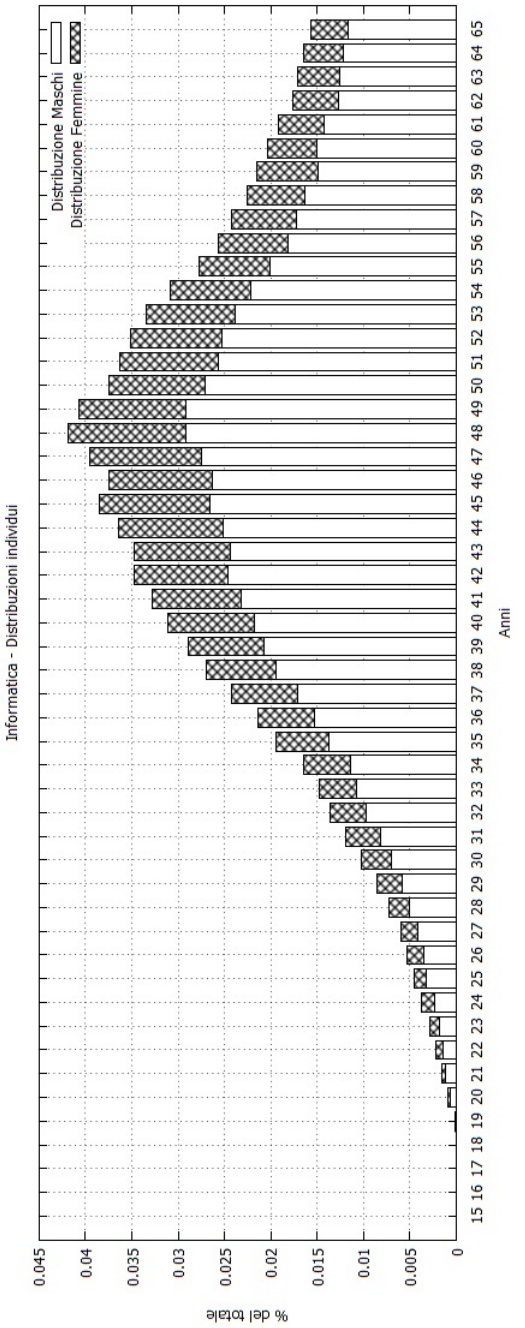


Figura 3.10: Distribuzione degli individui per fascia d'età - Rete informatica

φ	Archi	% del totale	Componenti connesse
1	5.495	71,42	8.449
2	1.043	13,56	10.213
3	488	6,34	10.590
4	211	2,74	10.790
≥ 5	457	5,94	10.739

Tabella 3.2: Archi nelle rete di aziende finanziarie

Figura 3.11: Rete di aziende finanziarie

Descrizione della popolazione

Gli amministratori analizzati in questo scenario sono stati 40.573 di cui il 80% sono dirigenti maschi, bisogna anche annotare che sono stati scartati 12.463 individui che rappresentano il 23.5% della popolazione complessiva, anche qui come per il caso precedente valgono le stesse motivazioni.

Nella figura 3.12 si osserva come la distribuzione degli individui per età vada ad accostarsi ad una gaussiana.

Si evidenzia come la differenza fra maschi e femmine sia ancora più accentuata rispetto alla rete informatica.

3.8 Scelta degli indici utilizzati

Gli indici utilizzati per misurare la segregazione sono stati l'indice di isolamento e l'entropia, rispettivamente descritti nella sezione 1.4.2 e 1.4.3.

É stata scelta la loro forma *binaria* perchè più vicina alla definizione di segregazione così come è stata data. Infatti si deve misurare quanto un gruppo si sia allontanato dalla comunità globale e la forma binaria risulta a nostro parere la più appropriata.

Entrambi gli indici sono espressi in funzione di un gruppo protetto e variano nell'intervallo [0;1].

L'indice di isolamento mostra quanto un gruppo protetto sia esposto a se stesso e questo offre una misura chiara di come tale gruppo sia distribuito all'interno delle unità organizzative. Più l'indice si avvicina ad 1 e maggiore è l'isolamento del gruppo dal resto della società, poiché tende a restare in unità organizzative in maggioranza frequentate da individui appartenenti al suo stesso gruppo.

L'entropia nella sostanza coincide con l'indice di isolamento, può essere vista come una misura che indica il grado di purezza delle unità organizzative, ovvero se la distribuzione di individui tende nettamente a favore del gruppo protetto o a favore del resto della comunità. In una unità organizzativa in cui il 90% degli individui appartiene ad un gruppo protetto, l'indice avrà un basso valore poiché vi è poca incertezza. Tale incertezza può essere vista come un indicatore che più si avvicina ad uno e più esprime equità nella distribuzione del gruppo protetto, poiché indica la difficoltà ad individuare minoranze all'interno di una unità organizzativa.

Un basso valore dell'entropia è indice di una ottima integrazione fra i gruppi protetti visto che la distanza sociale fra la comunità globale e le varie minoranze è minimizzata.

Il motivo per cui sono state utilizzate due misure differenti nel calcolo ma simili nella sostanza è per capire se vi è una corrispondenza nei risultati dati dalle due misurazioni.

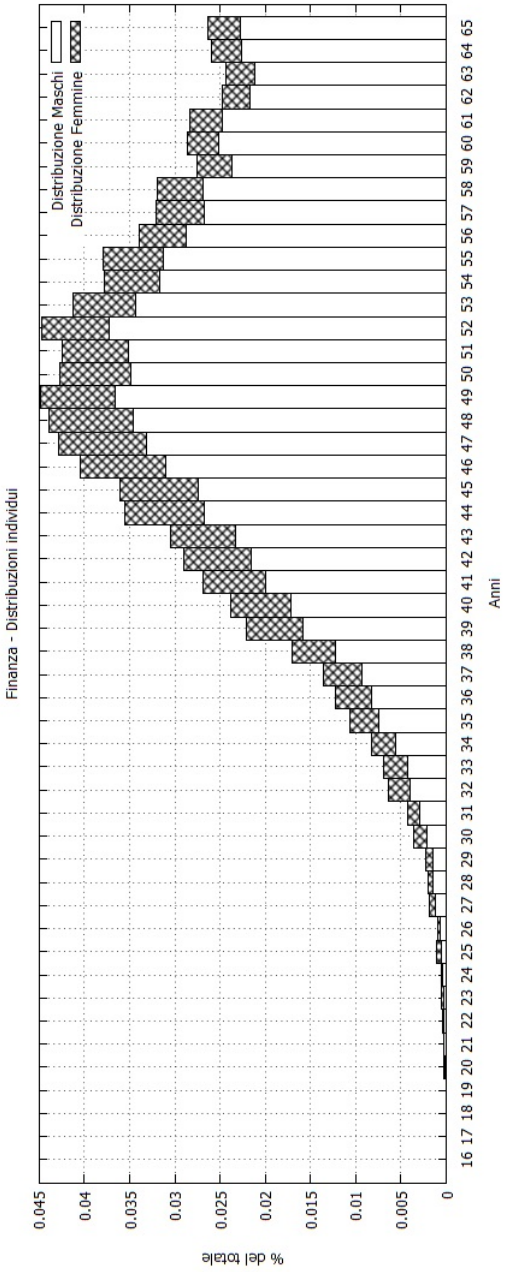


Figura 3.12: Distribuzione degli individui per fascia d'età - Rete finanza

3.9 Architettura del sistema

La struttura del sistema è stata definita da vari componenti che implementano i requisiti definiti in precedenza:

- gestore del processo *ETL*,
- gestore della rete sociale intesa come grafo non orientato,
- gestore del calcolo dell'indice di isolamento,
- gestore del calcolo dell'entropia.

Ovviamente tutte queste componenti sono controllate da un modulo che gestisce le varie fasi del calcolo.

Il diagramma delle classi nella figura 3.13 mostra in modo conciso le principali componenti del sistema come siano in relazione fra loro.

Di seguito si fornisce una breve descrizione per ognuna di esse

Main configura il comportamento del programma caricandolo da file.

Parser gestore del processo ETL[4].

GestoreDelGrafo costruisce e gestisce il grafo, calcola le comunità.

StrutturaDati memorizza le entità della rete sociale.

GestoreEntropia calcola l'entropia e scrive i risultati su file.

GestoreIndiceIsolamento calcola l' isolamento e scrive i risultati su file.

Amministratore rappresenta l'entità amministratore.

Azienda rappresenta l'entità azienda.

Il diagramma di sequenza in figura 3.14 mostra come i moduli interagiscano. Per prima cosa il *main* ordina al *GestoreDelGrafo* di creare il grafo ed in seguito di calcolare le comunità attraverso il procedimento descritto nella sezione 3.7. Quando il *main* possiede le unità organizzative può creare ed invocare i gestori degli indici. Non sono mostrati i moduli *amministratore e azienda* perchè semplici strutture dati che servono al *Parser* quando estrae i dati da file e li trasforma, per poi invocare le funzionalità del modulo *StrutturaDati* per caricarli nel sistema.

[4] Extract, Transform, Load

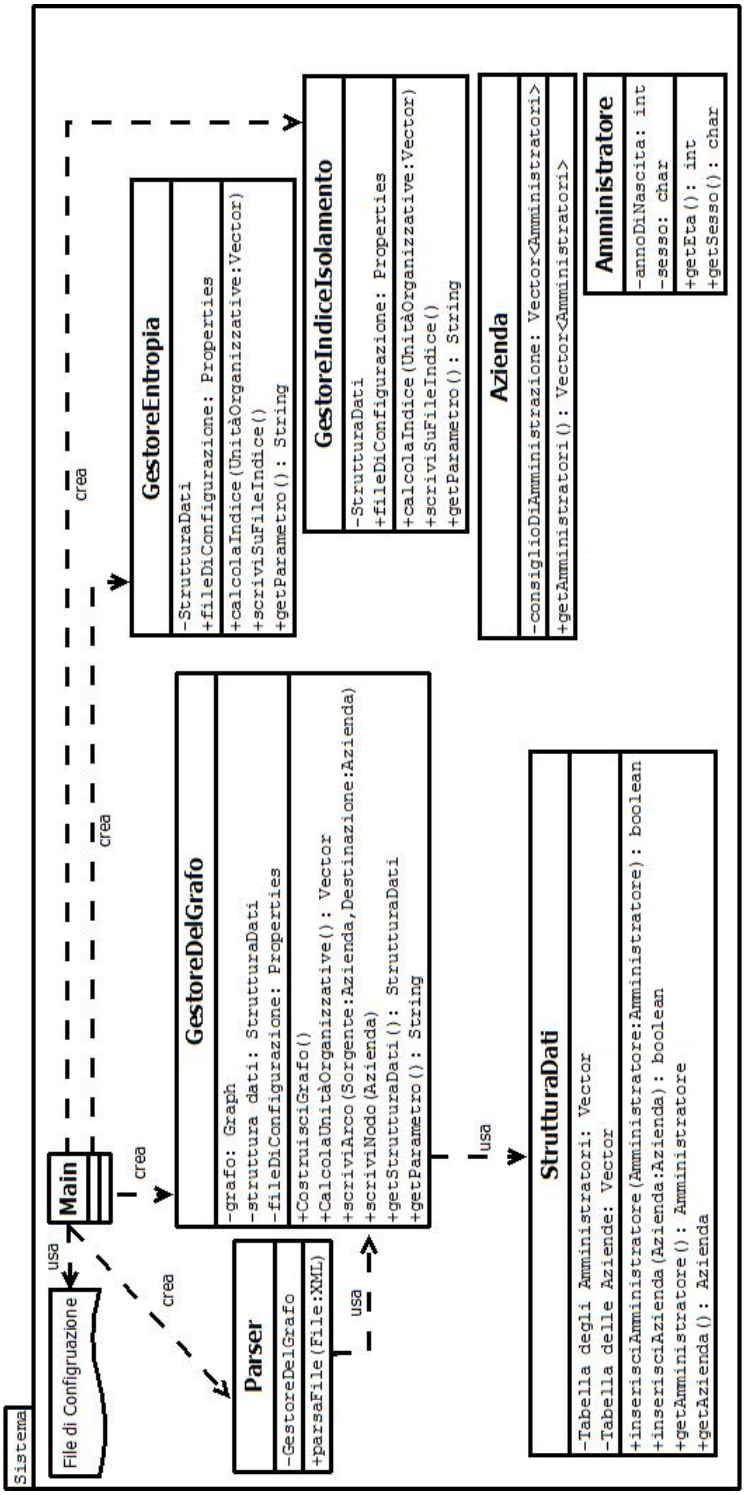


Figura 3.13: Diagramma delle classi del sistema

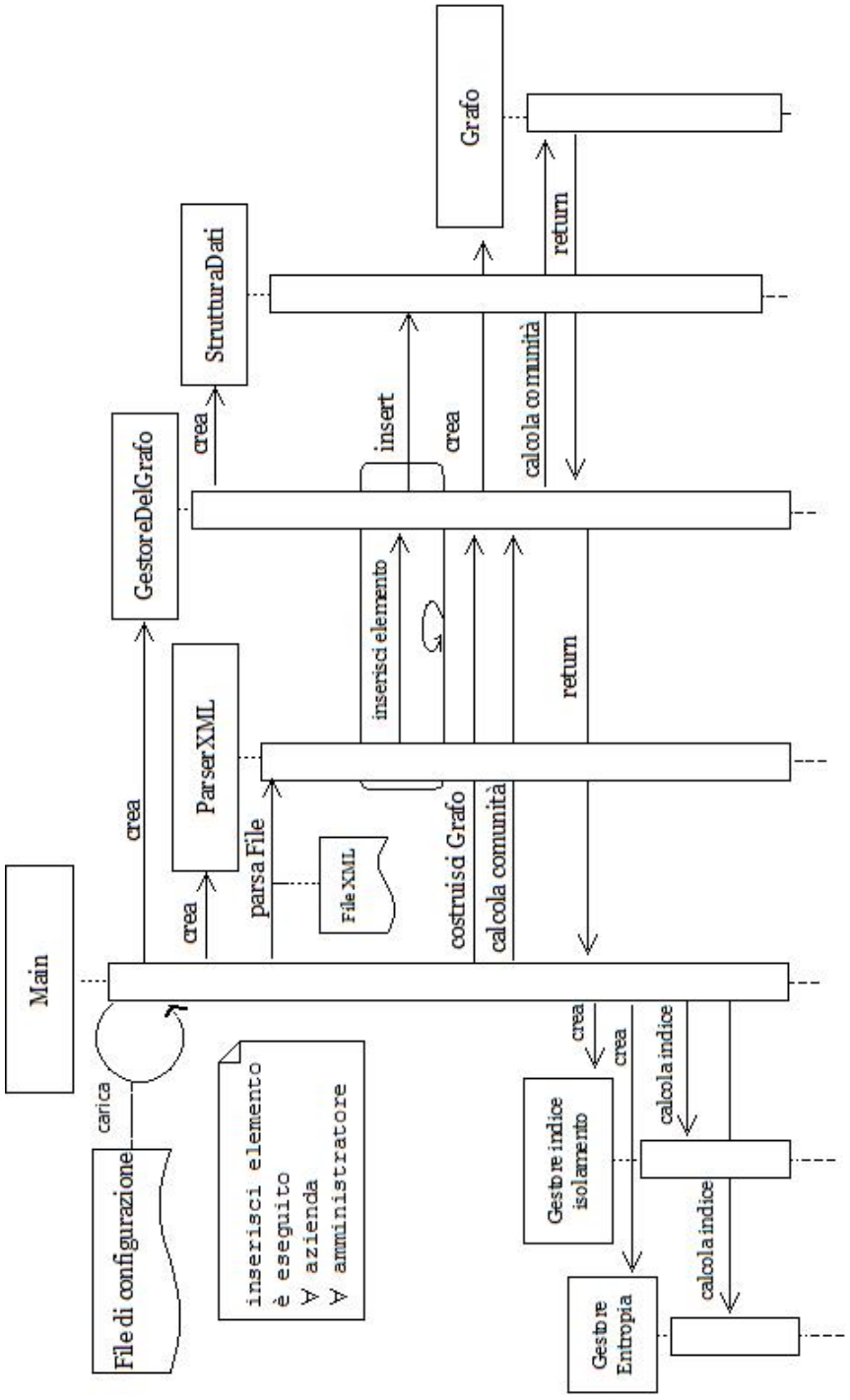


Figura 3.14: Diagramma di sequenza del sistema

Capitolo 4

Aspetti implementativi

4.1 Dettagli software

Il programma è stato scritto nel linguaggio di programmazione orientato agli oggetti *Java* versione SE 7 Edition. Come ambiente di sviluppo è stata utilizzata la piattaforma *Eclipse* versione Juno Service Release 1.

Librerie Esterne

Il progetto sfrutta la libreria *Jung* (Java Universal Network/Graph Framework) versione 2.0.1 per poter estrarre dal grafo le unità organizzative.

Il requisito di calcolare all'interno del grafo le unità organizzative è stato quindi affidato interamente a tale libreria e la descrizione della sua implementazione non sarà trattata nella tesi. Il 90% del tempo di esecuzione del sistema è legato all'utilizzo di questa libreria.

Si è utilizzata la API[1] *SAX 2.0.1* per il processo ETL.

4.2 Processo ETL

Il processo *ETL* è descritto in pseudo-codice simil Java nell' algoritmo 1 ed è la descrizione del metodo `parsaFile(File:XML)` del modulo Parser nella figura 3.13.

[1] Application Programming Interface

Algorithm 1 Processo ETL

Requisiti: A) Un *file XML* con uno schema descritto nella sezione 4.2.1.
B) Un'istanza non nulla del gestore del grafo z.

Restituisce: VOID - Modifica all'interno del gestore del grafo, la struttura dati contenente tutte le aziende e gli amministratori.

```
1:  etàMax ← getParametro(etàMax)
2:  etàMin ← getParametro(etàMin)
3:  pathFileXML ← getParametro(pathFile)
4:  fileXML ← apriFile(pathFileXML)
5:  while fileXML ha elementi do
6:      elemento ← fileXML.leggiProssimoElemento()
7:      if elemento è di tipo Azienda then
8:          z.getStrutturaDati().scriviAzienda(elemento)
9:          for all amministratore z ∈ elemento do
10:             if z.getEta≥etàMin && z.getEta≤ etàMax then
11:                 z.getStrutturaDati().scriviAmministratore(z)
12:             end if
13:         end for
14:     end if
15: end while
```

4.2.1 Formato del file di input

I dati da cui è ricavata la rete sono stati forniti da "InfoCamere"[2] una società consortile di informatica delle Camere di Commercio Italiane per azioni e sono memorizzati in un file scritto nel linguaggio di *markup* chiamato *xml* [3].

I dati rappresentano le aziende con i loro dettagli identificativi, gli amministratori al comando delle stesse.

Sono stati forniti due file che descrivono i dati delle aziende appartenenti a due settori differenti:

- finanziario,
- informatico.

Nei file *xml* esiste un elemento root che indica il settore aziendale

`<informatica>` oppure `<finanza>`

Al suo interno contiene un insieme di imprese le cui informazioni sono contenute in due tag `<DATI_IDENTIFICATIVI>` e `<DATI_IMPRESA>`, come vede dalla figura 4.1. All'interno del secondo tag sono contenuti tutti gli amministratori dell'impresa con le relative informazioni.

[2] è la struttura tecnologica di eccellenza a supporto del patrimonio informativo e di servizi del sistema camerale.

[3] *eXtensible Markup Language*

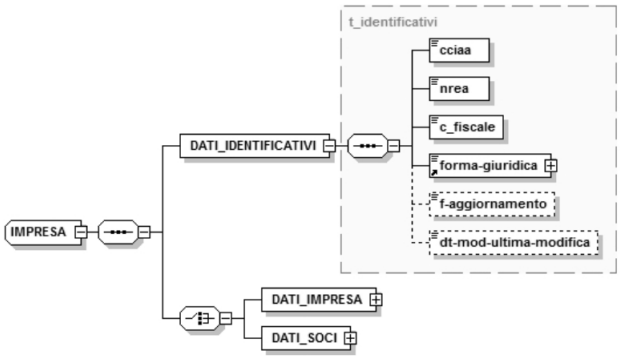


Figura 4.1: Componente Impresa

4.3 Creazione della rete sociale

Utilizzando la struttura dati creata dal *GestoreDelGrafo* e riempita dal *Parser* attraverso il processo *ETL*, viene creata una rete in cui i nodi sono le aziende mentre un arco fra una azienda A ed una azienda B viene instaurato se le due aziende hanno almeno un numero parametrico φ (caricato dal file di configurazione) di dirigenti in comune.

Il processo è descritto nello pseudo-codice dell'algoritmo 2 ed è la descrizione del metodo `costruisciGrafo()` del modulo GestoreDelGrafo nella figura 3.13.

Algorithm 2 Creazione del grafo

Requisiti: Istanza non nulla della struttura dati κ

Restituisce: VOID - Modifica l'oggetto grafo ϑ all'interno del *GestoreDelGrafo*, inserendo tutte le aziende con almeno φ amministratori comuni.

1: $\varphi \leftarrow getParametro(\varphi)$
2: **for all** azienda $i \in \kappa$ **do**
3: scriviNodo(i)
4: $amministratori_i = i.getAmministratori()$
5: **for all** azienda $j \in k \mid (j \neq i \;\&\&\; j$ non visitata$)$ **do**
6: $amministratori_j = j.getAmministratori()$
7: $pesoArco \leftarrow amministratori_j \cap amministratori_i$
8: **if** $pesoArco \geq \varphi$ **then**
9: $scriviArco(i, j)$
10: **end if**
11: **end for**
12: segna i come azienda visitata
13: **end for**

4.4 Calcolo dell'indice di isolamento

4.4.1 Misura effettiva

Una misura utilizzata per scoprire segregazione è l'indice di isolamento, si veda la sezione 1.4.2. Partendo dall'equazione 4.1, in cui GP_i sta ad indicare la dimensione del gruppo protetto appartenente all'unità organizzativa i, dim_{GP} indica la dimensione nella rete del gruppo protetto studiato, dim_i indica la dimensione complessiva dell'unità organizzativa i.

$$\gamma = \sum_i \left(\frac{GP_i}{dim_{GP}} \cdot \frac{GP_i}{dim_i} \right) = \underbrace{\frac{1}{dim_{GP}}}_{\boldsymbol{\mu}} \cdot \underbrace{\sum_i \frac{GP_i^2}{dim_i}}_{\boldsymbol{\omega}} \tag{4.1}$$

La procedura di calcolo del indice è spiegata in pseudo-codice nell'algoritmo 3, in ingresso richiede l'insieme delle unità organizzative su cui calcolare la misura e la struttura dati su cui sono mantenute le informazioni; l'algoritmo calcola il valore dell'indice per ogni gruppo protetto.

Nella decima riga dell'algoritmo 3 si controlla che l'individuo in questione non sia già stato classificato poiché all'interno di una unità organizzativa possono esistere due o più aziende con un amministratore in comune e quindi può presentarsi la possibilità di incorrere in un elemento già classificato.

Il metodo `setGPDimUnitaOrganizzativa(`j`)` aggiorna la struttura dati relativa al gruppo protetto di appartenenza dell'individuo j (gpiMaschi o gpiFemmine) si veda la figura 4.2.

Il metodo `aggiornaDatiUnita(` $dimensione_i$`)` utilizza le strutture dati contenenti le dimensioni dei gruppi protetti dell'unità i (gpiMaschi e gpiFemmine) e calcola il ramo ω dell'equazione 4.1. La sommatoria viene implementata sommando il risultato di tale calcolo alla struttura dati che contiene le misure dell'indice (isolamentoMaschi, isolamentoFemmine, isolamentoMisti).

Infine il metodo `aggiornaDatiGlobali` si occupa per ogni valore dell'età x e del sesso y, di calcolare $dim_{GP_{xy}}$. Ogni valore $dim_{GP_{xy}}$ è utilizzato come divisore sugli oggetti che contengono le misure dell'indice.

L'implementazione del calcolo dell'indice ha un costo di esecuzione approssimato per difetto a

$$\mathcal{O}(\underbrace{(U \cdot A \cdot D \cdot \nu)}_{\boldsymbol{\alpha}} + \underbrace{U \cdot (\xi)}_{\boldsymbol{\beta}} + \underbrace{\mathcal{O}(\alpha)}_{\boldsymbol{\epsilon}})$$

in cui U rappresenta il numero delle unità organizzative, A la dimensione massima di una unità organizzativa (data dal numero di aziende), D il numero massimo di dipendenti di una azienda, N_{GP} il numero di gruppi protetti di cui si vuole l'indice.

Ecco la spiegazione dei vari addendi:

$\boldsymbol{\alpha}$ si calcola per tutti gli amministratori a quale gruppo protetto appartengono, ν è il costo per calcolare `setGPDimUnitaOrganizzativa(`j`)`.

$\boldsymbol{\beta}$ si calcola la funzione `aggiornaDatiUnita(` $dimensione_i$`)` di costo ξ per ogni unità organizzativa.

$\boldsymbol{\epsilon}$ si calcola il ramo μ dell'equazione 4.1 attraverso l'utilizzo della funzione `aggiornaDatiGlobali` di costo α.

Algorithm 3 Calcolo dell'indice di isolamento γ

Requisiti: Dati

A) Un insieme di unità organizzative β calcolate dal GestoreGrafo.

B) Un oggetto StrutturaDati κ, copia di quello nel GestoreDelGrafo.

Restituisce: VOID - Assegna ai 3 oggetti che rappresentano il calcolo dell'indice per il sesso, le misure calcolate.

```
1:  for all i unità organizzativa ∈ β do
2:      lavoratoriVisitati ← new set()
3:      gpiMaschi ← new Vector(N_GP)
4:      gpiFemmine ← new Vector(N_GP)
5:      dimensione_i ← 0
6:      for all azienda z ∈ i do
7:          z ← k.getAzienda()
8:          for all dipendente j ∈ z do
9:              j ← k.getAmministratore()
10:             if j ∉ lavoratoriVisitati then
11:                 setGPDimUnitaOrganizzativa(j)
12:                 lavoratoriVisitati.inserisci(j)
13:             end if
14:         end for
15:     end for
16:     aggiornaDatiUnita( dimensione_i)
17: end for
18: aggiornaDatiGlobali()
19: return γ
```

4.4.2 Misura ideale per confronto

É stato calcolata anche una versione "ideale" dell'indice, affinché fosse possibile confrontare i valori nello scenario reale con quelli in un mondo ideale, privo di segregazione così come è definita nella sezione 3.3.

La formula 4.1 mostra come l'indice non sia altro che la sommatoria per ogni unità organizzativa i del prodotto di due quote:

- la percentuale di individui del gruppo protetto $\in i$ $(\frac{GP_i}{dim_{GP}})$

- la percentuale di individui di $i \in$ al gruppo protetto $(\frac{GP_i}{dim_i})$.

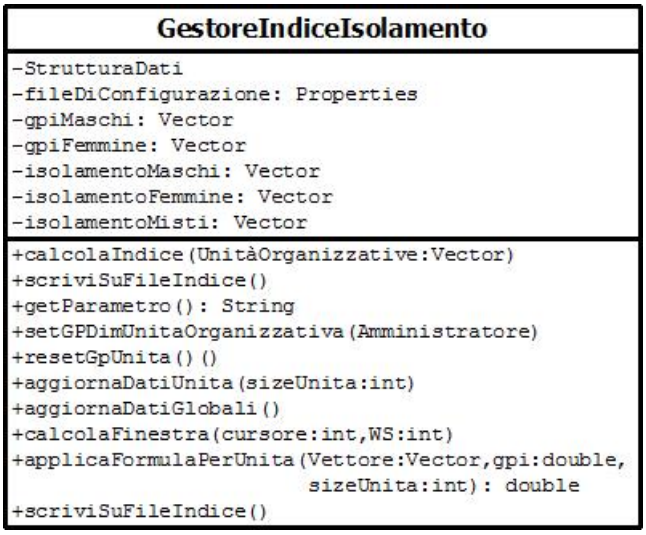


Figura 4.2: Dettaglio della classe GestoreIndiceIsolamento

Si è tratto ispirazione dal pensiero di Mark Fosset [10] per cui l'integrazione viene vista come la possibilità di ritrovare all'interno dei quartieri una distribuzione degli individui che rispetti la distribuzione globale della popolazione. Questo significa che se il 20% della popolazione globale appartiene al gruppo protetto x, nello scenario ideale di integrazione ogni unità organizzativa dovrebbe possedere la stessa proporzione del 20% di individui appartenenti al gruppo x. Per questo motivo i due fattori descritti sopra che componevano il prodotto vanno esattamente a coincidere.

Nella formula 4.1 viene sommato il contributo di ogni unità organizzativa al valore complessivo dell'indice di isolamento, poiché ovviamente ogni unità possiede percentuali differenti del gruppo protetto, nel caso ideale invece ogni unità presenta la stessa realtà, ecco quindi che $\gamma = \sum_i \left(\frac{GP_i}{dim_{GP}} \cdot \frac{GP_i}{dim_i}\right)$ diventa

$$\gamma = \sum_i \left(\frac{GP}{dim_{tot}} \cdot \frac{GP}{dim_{tot}}\right) = n \cdot \left(\frac{1}{n} \cdot \frac{GP}{dim_{tot}} \cdot \frac{GP}{dim_{tot}}\right) = \left(\frac{GP}{dim_{tot}}\right)^2 \quad (4.2)$$

in cui GP indica la dimensione del gruppo protetto in questione e dim_{tot} indica il numero totale di individui classificati.

4.5 Calcolo dell'entropia

Questo tipo di misura, se ne accenna nella sezione 1.4.3, trova un suo frequente utilizzo come misuratore di purezza dei classificatori nel data mining.

$$\eta = -\sum_{i}^{n} p(a_i) \cdot \log(p(a_i)) \tag{4.3}$$

Partendo dall'equazione 4.3, tale misura può essere vista come il grado di incertezza in merito ad un numero finito di alternative (a_i, ..., a_n) non equiprobabili, in cui p(a_i) indica la probabilità che si verifichi l'evento a_i.

Nel contesto della segregazione, un evento p(a_i) può essere interpretato come la frequenza di un gruppo protetto all'interno di una unità organizzativa.

L'entropia è calcolata come una media ponderata dal peso P_i di ogni unità organizzativa i. Tale peso è definito come la quota del gruppo protetto all'interno di ogni unità i.

La forma binaria è definita nel seguente modo

$$\eta = -\sum_{i} P_i \cdot \left(\frac{GP_i}{dim_i} \log \left(\frac{GP_i}{dim_i} \right) + \frac{dim_i - GP_i}{dim_i} \log \left(\frac{dim_i - GP_i}{dim_i} \right) \right) \tag{4.4}$$

Nell'equazione 4.4 si somma per ogni unità organizzativa i la quota di entropia di tale unità ponderata per il suo peso definito come $P_i = \frac{dim_i}{\beta}$.

Viene espressa l'entropia per ogni comunità in cui un individuo o appartiene al gruppo protetto GP_i oppure al resto della comunità, definito come $dim_i - GP_i$.

Non sono riportati né lo pseudo-codice per calcolare l'entropia né i dettagli della classe gestoreEntropia perchè uguali alla classe gestoreIndiceIsolamento. L'unica differenza è nell'applicazione della formula che descrive le due misure.

Capitolo 5

Analisi dei risultati

Il software è stato eseguito passando come input entrambi i file relativi alle reti di informatica e finanza.

Sono stati utilizzati i seguenti parametri:

- φ di cui si parla nella sezione 3.6.2 è stato impostato a 1, 2, 3.
- WS di cui si parla alla sezione 3.4.1, è stato impostato a 3.

Sono stati utilizzati tutti gli approcci definiti nella sezione 3.4

- Approccio a taglio ottimo binario - Modo 1
- Approccio a taglio ottimo binario - Modo 2
- Approccio media mobile - Modo 3

I grafici illustrano sull'asse delle x l'età mentre sull'asse delle y l'indice di segregazione in funzione del gruppo GP_x.

In questo modo, per ogni possibile gruppo protetto, viene espresso il relativo valore dell'indice in modo da esprimere possibili correlazioni fra i valori dell'attributo età.

Le considerazioni che valgono per tutti i grafici analizzati sono le seguenti:

– Per i gruppi "maturi" (linea blu e nera) individuati con il *modo 1*, si osserva la parte destra del grafico, *la più significativa per questo tipo di metodo*, perchè spostarsi a sinistra del grafico equivale a collassare i vincoli sull'età e quindi a ritenere tutti gli individui della rete appartenenti al gruppo protetto (senza considerare il sesso).

– Per i gruppi "giovani" (linea rossa e marrone), individuati con il *modo 2*, si osserva la parte sinistra del grafico, *la più significativa per questo tipo di metodo*, per la situazione inversa vista in precedenza.

  Peresempio $GP_{15_{maschi}}$ individuato con il *modo 1* equivale a $GP_{65_{maschi}}$ con il *modo 2* e rappresentano l'indice in funzione dei maschi.

- Il variare di φ non comporta alcun cambiamento nei risultati .

  Si ipotizza a causa delle caratteristiche della rete (la rete è fortemente sconnessa) per cui non fa differenza l'eliminazione dei legami più deboli.

  Di conseguenza tutti i grafici utilizzati sono espressione dell'esecuzione del sistema per la rete ottenuta con $\varphi = 1$.

  Le figure 5.3 e 5.4 sono le uniche eccezioni (illustrano l'esecuzione con $\varphi = 3$) e sono riportate per confermare che l'esecuzione variando φ risulta equivalente (le corrispettive figure 5.1 e 5.2 con $\varphi = 1$).

- Una differenza sostanziale fra l'indice reale e quello reale.

  Si ipotizza poiché le unità organizzative sono fortemente composte da gruppi omogenei di individui. In una realtà "*equa*" questo non dovrebbe accadere.

Le sezioni successive mostreranno i risultati commentati suddivisi in base alla specifica rete.

5.1 Rete informatica

5.1.1 Indice di isolamento

Gruppi misti

La figura 5.1 mette a confronto i gruppi "*giovani*" con i "*maturi*" mentre la figura 5.2 offre una istantanea dell'approccio della media mobile.

Le considerazione oggettive che possono essere fatte sono le seguenti:

1. Per i gruppi "maturi" l'inclinazione della linea del'indice ideale (matematicamente parlando la sua derivata) aumenta repentinamente nell'intervallo delle x [40;50] questo è naturale poiché nella figura 3.10 si può osservare che il peso di quella fascia d'età è notevole e ciò porta alla naturale conseguenza dell'aumento dell'indice. Più aumenta la quota globale di persone che consideriamo appartenenti ad un gruppo protetto e più questo indice aumenta, come si può vedere nella definizione nella sezione 4.4.

2. Per i gruppi "giovani" si osserva come l'isolamento si innalzi repentinamente a differenza della misura ideale, questo indica una forte segregazione per la fascia d'età [16;20]. Un possibile motivo risiede nella tipica gestione familiare delle aziende e dalla frequente abitudine di inserire come figure fittizie i figli all'interno dei consigli di amministrazione.

3. Per l'approccio media mobile si osserva come la segregazione cresca fino ai 30 anni, tale soglia una volta raggiunta rappresenta il massimo assoluto, di fatti al suo superamento la segregazione resta stabile fino ai 50 anni.

Nella figura 5.5 e 5.6 sono stati rappresentati due modi differenti di vedere la correlazione fra l'indice reale ed ideale. La prima figura esprime per ogni età la percentuale di differenza fra l'indice reale e quello ideale: i GP_{20} sono distanti dalla realtà al 75%. La seconda indica il rapporto della misura reale su quella ideale. Non sono stati utilizzati per le analisi perchè a nostro parere poco intuitivi.

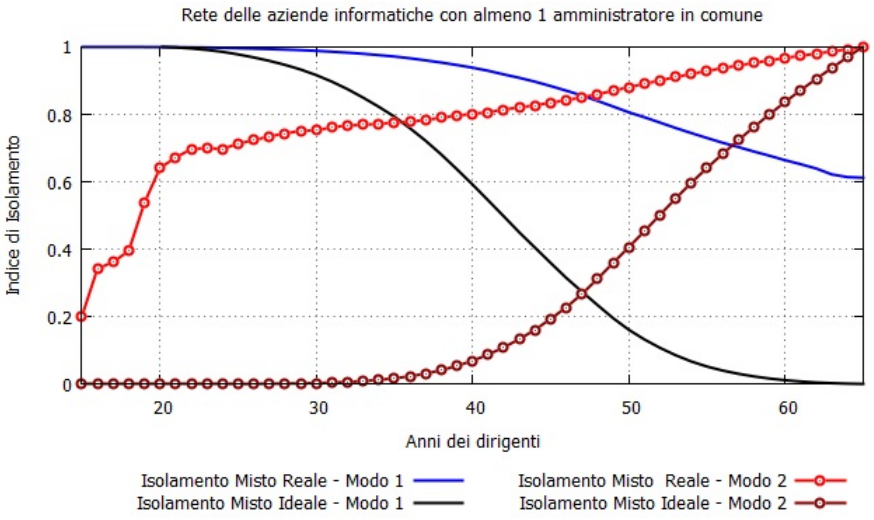


Figura 5.1: Isolamento - Informatica - Misto - Modo 1 e 2

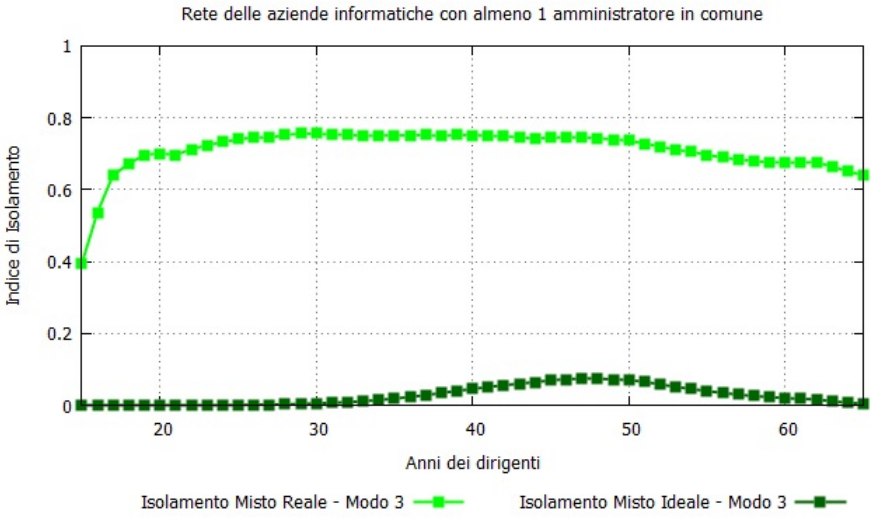


Figura 5.2: Indice di isolamento - Informatica - Misto - Modo 3

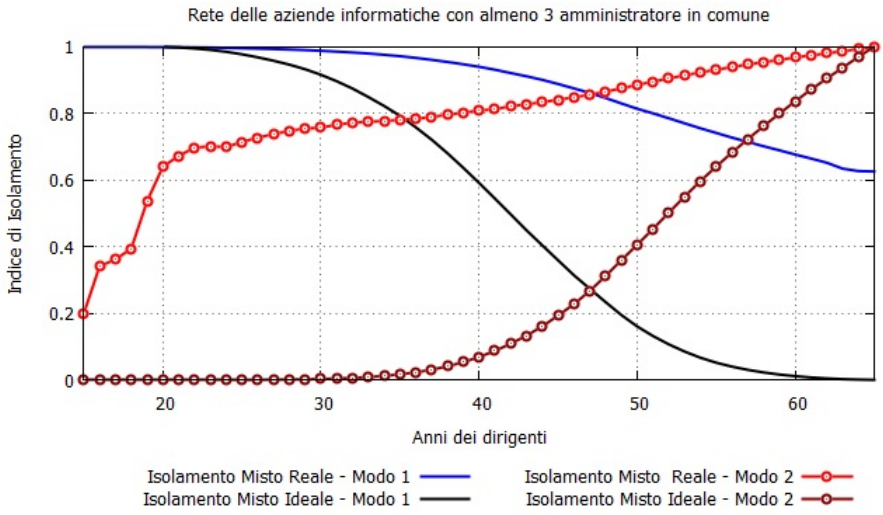


Figura 5.3: Isolamento - Informatica - Misto - Modo 1 e 2 - $\varphi = 3$

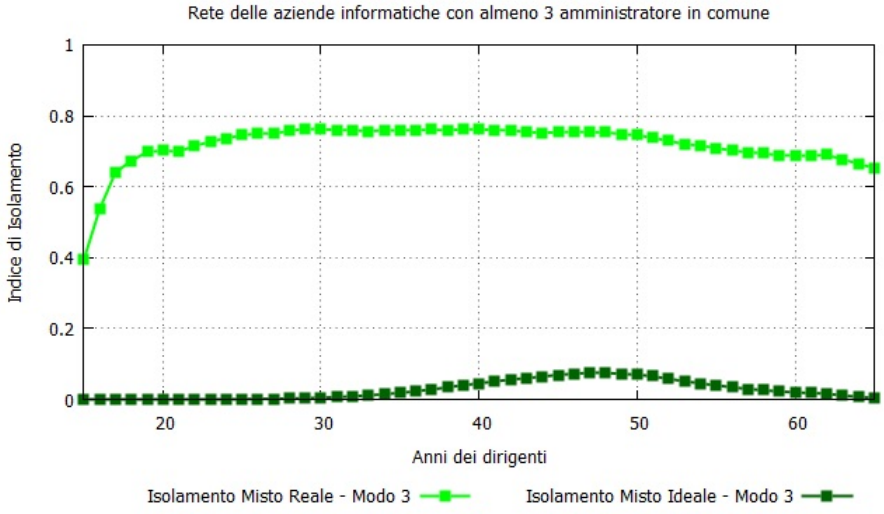


Figura 5.4: Isolamento - Informatica - Misto - Modo 3 - $\varphi = 3$

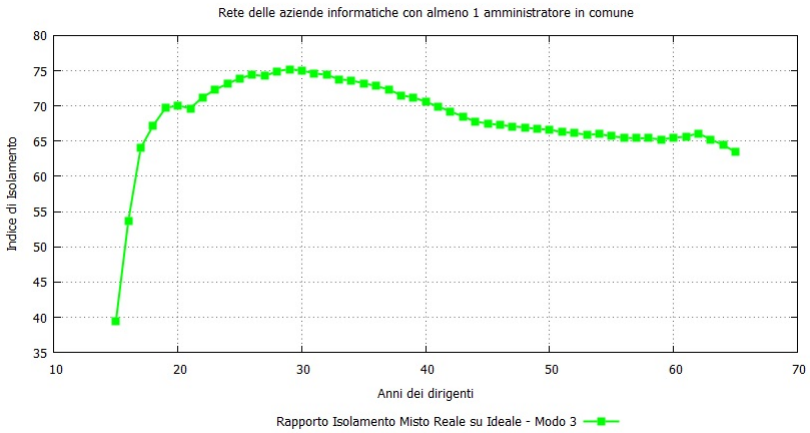


Figura 5.5: Isolamento - Informatica - Misto - Modo 3 - (Reale-Ideale) in %

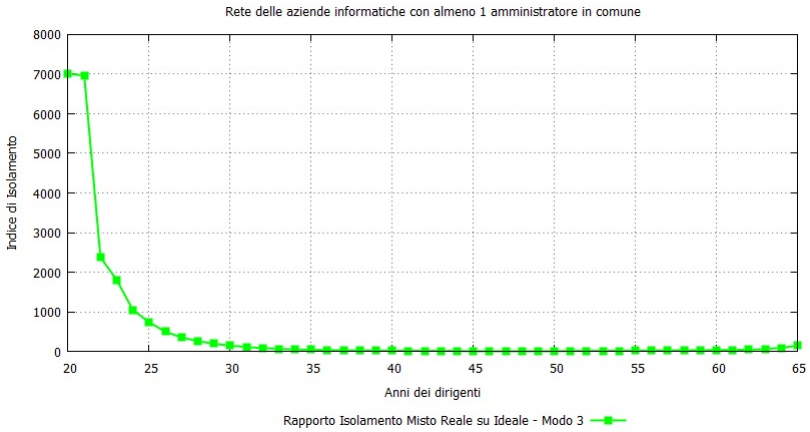


Figura 5.6: Isolamento - Informatica - Misto - Modo 3 - $\frac{Reale}{Ideale}$

Gruppi maschili

Nelle figure 5.7 e 5.8 si mostrano i risultati in cui si considerano gli amministratori di sesso maschile.

Le considerazione oggettive che possono essere fatte sono le seguenti:

1. Per i gruppi "maturi" individuati con il *modo 1* (linea blu e nera) si riesce a trovare un movimento di linea molto simile all'andamento che dovrebbe essere quello ideale, questo ad indicare che l'apporto che fornisce ogni età aggiunta al gruppo protetto è uguale a quello dell'indicatore ideale e questo indica che probabilmente l'età in questo gruppo non è una discriminante.

2. Per i gruppi "giovani" individuati con il *modo 2* (linea rossa e marrone) si nota come siano più segregati rispetti ai più anziani.

   Ad esempio il gruppo dei 22enni GP_{22} (linea rossa) paragonato al gruppo dei 60enni GP_{60} (linea blu) risulta più segregato. Inoltre ogni incremento d'età, nel gruppo dei giovani, fa aumentare significativamente il valore reale dell'indice e ciò a conferma di quanto detto prima.

3. Per i gruppi individuati con la media mobile (linea verde e verde scuro) si nota come per le età nell'intervallo [15;20] vi sia una forte variazione nelle misure, ad indicare una anomalia nella segregazione della fascia giovanile. La segregazione diminuisce mano a mano che l'età aumenta, partendo dai 50 anni. Questo conferma che gli amministratori più "maturi" sono proporzionalmente equidistribuiti fra le unità.

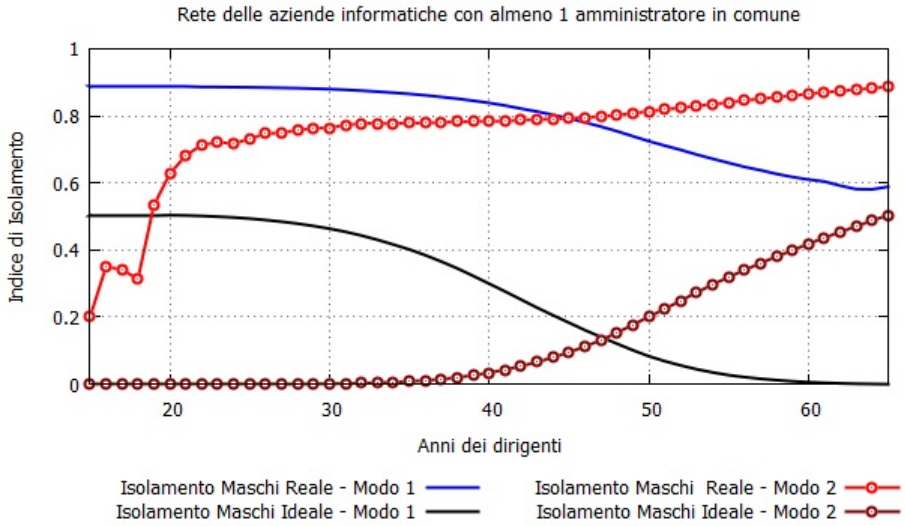


Figura 5.7: Isolamento - Informatica - Maschi - Modo 1 e 2

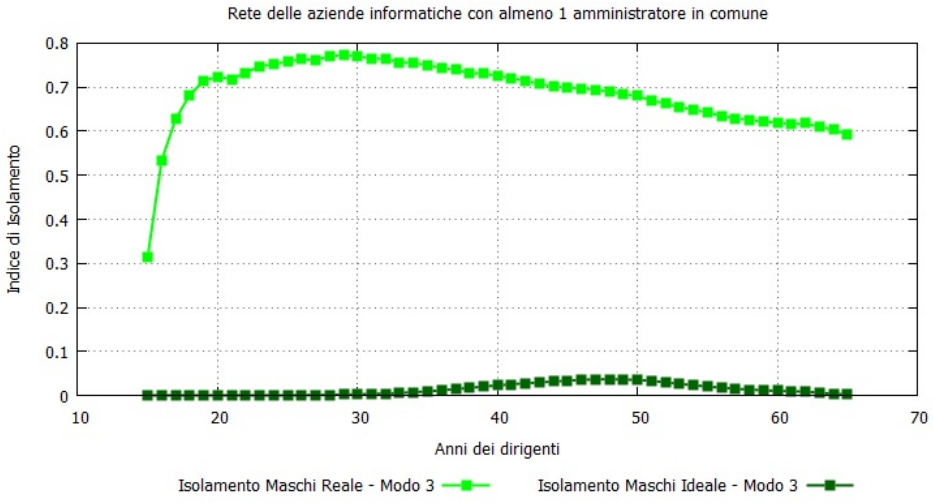


Figura 5.8: Isolamento - Informatica - Maschi - Modo 3

Gruppi femminili

Nelle figure 5.9 e 5.10 si mostrano i risultati in cui si considerano gli amministratori di sesso femminile.

Le considerazione oggettive che possono essere fatte sono le seguenti:

1. Per i gruppi "maturi" si nota come la curva reale non abbia un crescendo come negli altri casi ma al contrario presenti una convessità nell'intervallo [50;60] anni che indica come per quelle età l'indice reale in proporzione diminuisca. Questo indica una concentrazione di femmine in unità di grandi dimensioni a discapito della distribuzione equa che si dovrebbe avere.

2. I gruppi "giovani" sono discriminati come per i maschi.

3. Per i gruppi femminili individuati con la media mobile (linea verde e verde scuro), l'andamento dell'indice reale è costante: l'età non è una discriminante.

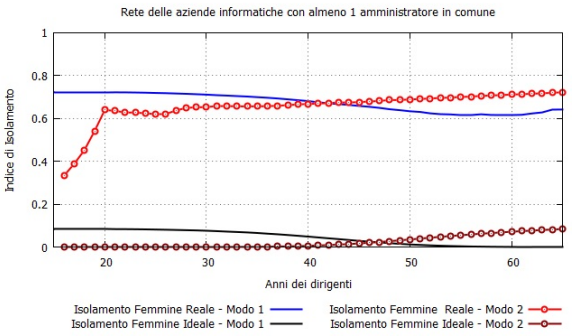


Figura 5.9: Isolamento - Informatica - Femmine - Modo 1 e 2

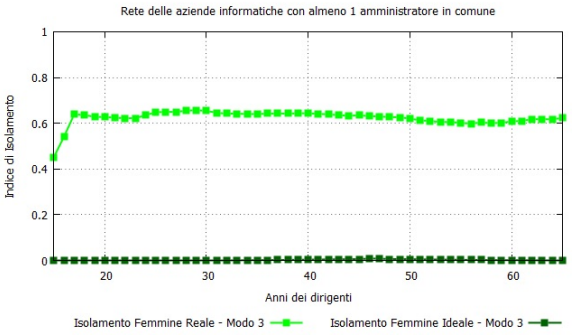


Figura 5.10: Isolamento - Informatica - Femmine - Modo 3

5.1.2 Entropia

Gruppi misti

Nella figura 5.11 sono mostrati i risultati ignorando il sesso.

Le considerazione oggettive che possono essere fatte sono le seguenti:

1. La tipica forma a campana gaussiana mostra come l'entropia raggiunga il suo massimo per gli anni nell'intervallo [45;50] e questo indica che per quella fascia d'età in proporzione vi è meno segregazione, poiché gli individui sono meglio distribuiti.

2. Le estremità della gaussiana negli intervalli [15:23] e [58;63] presentano due inclinazioni diverse, la prima più schiacciata rispetto alla seconda. Questo indica che i giovani sono più segretati in confronto agli amministratori anziani.

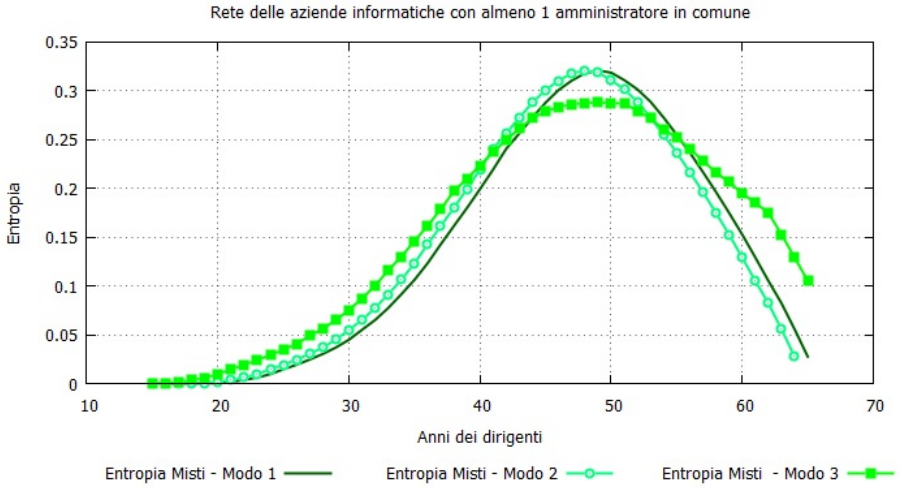


Figura 5.11: Entropia - Informatica - Gruppo Misti

Gruppi maschili

Nella figura 5.12 sono mostrati i risultati con tutte le modalità di individuazione in cui si considerano gli amministratori di sesso maschili.

Le considerazione oggettive che possono essere fatte sono le seguenti:

1. I gruppi "maturi" (linea blu), nella parte destra del grafico, sono segregati come già detto per il gruppo misto, ma in maniera inferiore al gruppo dei giovani (come già verificato con l'utilizzo dell'indice di isolamento).

2. Il massimo assoluto è inferiore al gruppo misto questo indica che è più probabile trovare consigli di amministrazione in cui ci siano tutti uomini, mentre le donne avendo un indice più basso sono distribuite in riferimento ai maschi più uniformemente, dato riscontrato anche nell'indice di isolamento.

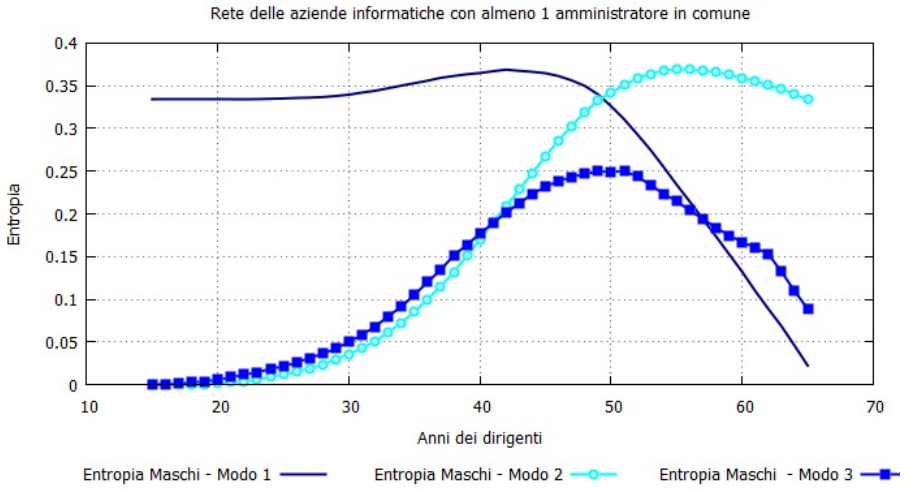


Figura 5.12: Entropia - Informatica - Gruppo Maschili

Gruppi femminili

Nella figura 5.13 sono mostrati i risultati con tutte le modalità di individuazione in cui si considerano gli amministratori di sesso femminili.

Le considerazione oggettive che possono essere fatte sono le seguenti:

1. La campana è piatta e ciò indica come già visto che l'età apparentemente non sia una discriminante.

2. Le femmine fra i [30;40] anni e fra i [55;65] sono più segregate degli uomini della stessa fascia.

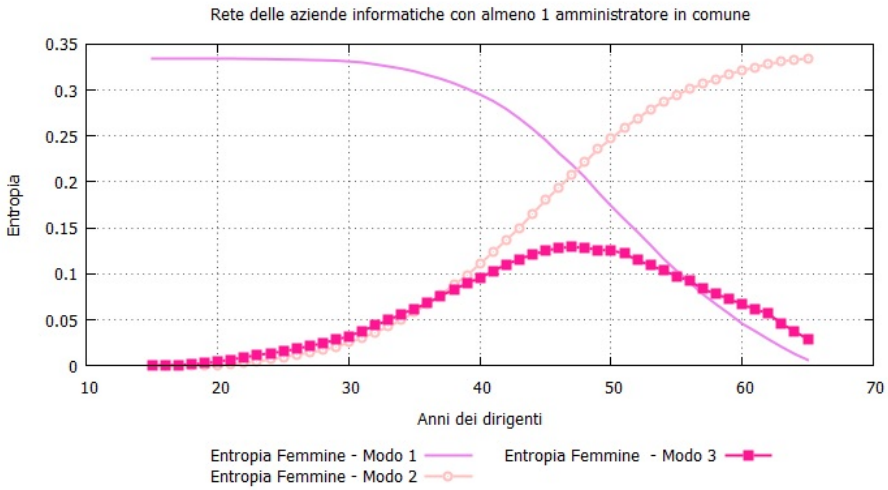


Figura 5.13: Entropia - Informatica - Gruppo Femmine

5.2 Rete finanziaria

5.2.1 Indice di isolamento

Gruppi misti

La figura 5.14 mette a confronto i gruppi "*giovani*" con i "*maturi*" mentre la figura 5.15 offre una istantanea dell'approccio della media mobile.

Le considerazione oggettive che possono essere fatte sono le seguenti:

1. Per i gruppi "maturi" (linea blu) la misura reale si avvicina maggiormente a quella ideale e questo vuol dire che sono in proporzione meno segregati che in informatica.

2. Per i gruppi "giovani" (linea rossa) sono meno segregati di quelli in informatica. Nell'intervallo [23;35] la segregazione cala fino a raggiunge gradualmente il suo minimo assoluto.

3. Per i gruppi individuati con la media mobile (linea verde chiaro) il minimo della segregazione viene raggiunto verso i 33 anni. Nell'intervallo [40;65] il comportamento è identico a quello ideale, con le debite proporzioni, l'età non è una discriminante. La fascia [15;20] è la più segregata ma potrebbe essere semplicemente per tornaconto familiari di proprietari aziendali in cui queste figure sono fittizie.

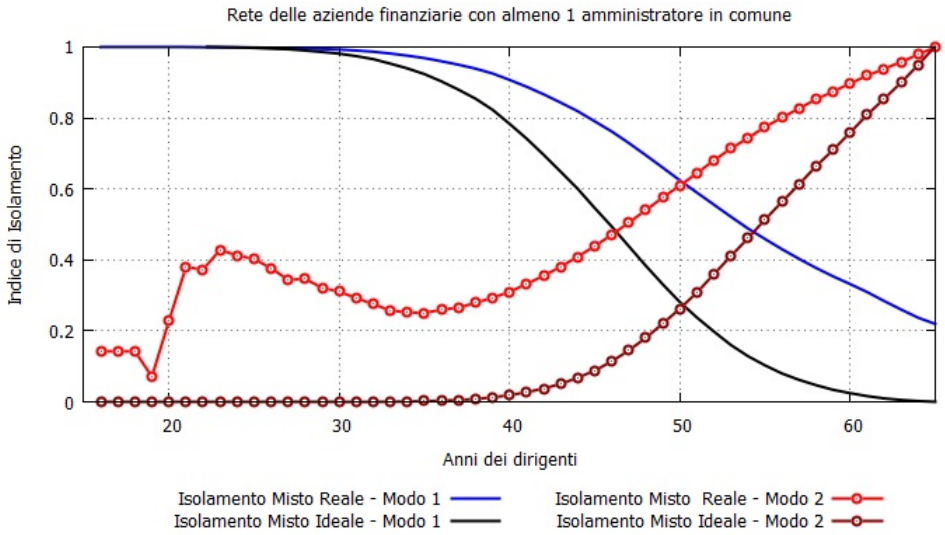


Figura 5.14: Isolamento - Finanza - Misto - Modo 1 e 2

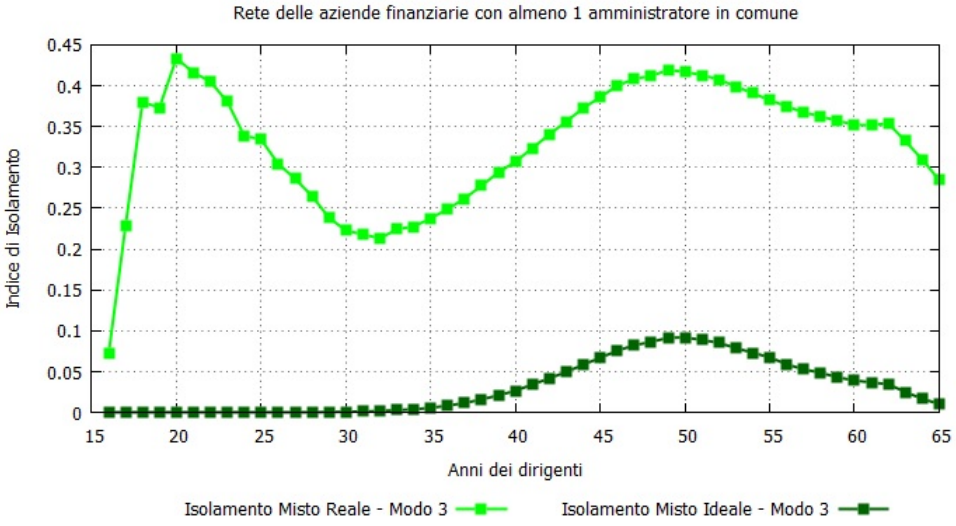


Figura 5.15: Isolamento - Finanza - Misto - Modo 3

Gruppi maschili

Nelle figure 5.16 e 5.17 si mostrano i risultati in cui si considerano gli amministratori di sesso maschile.

Le considerazione oggettive che possono essere fatte sono le seguenti:

1. Valgono le stesse considerazioni fatte per i gruppi misti.
2. I maschi presentano un valore dell'indice di isolamento di circa 0.8 come la rete informatica. La misura è indicata dalla linea blu nella parte sinistra del grafico o dalla linea rossa nella parte destra del grafico.

Gruppi femminili

Nelle figure 5.18, e 5.19 si mostrano i risultati in cui si considerano gli amministratori di sesso femminile.

Le considerazione oggettive che possono essere fatte sono le seguenti:

1. Per i gruppi "giovani" (linea rossa) il massimo viene raggiunto a 24 anni. Nell'intervallo (24;34] la segregazione cala gradualmente fino a raggiungere il suo minimo assoluto.
2. Per i gruppi individuati con la media mobile (linea verde chiaro) si osserva come nell'intervallo [33;65] la segregazione aumenti gradualmente, con un massimo relativo intorno ai 46 anni.

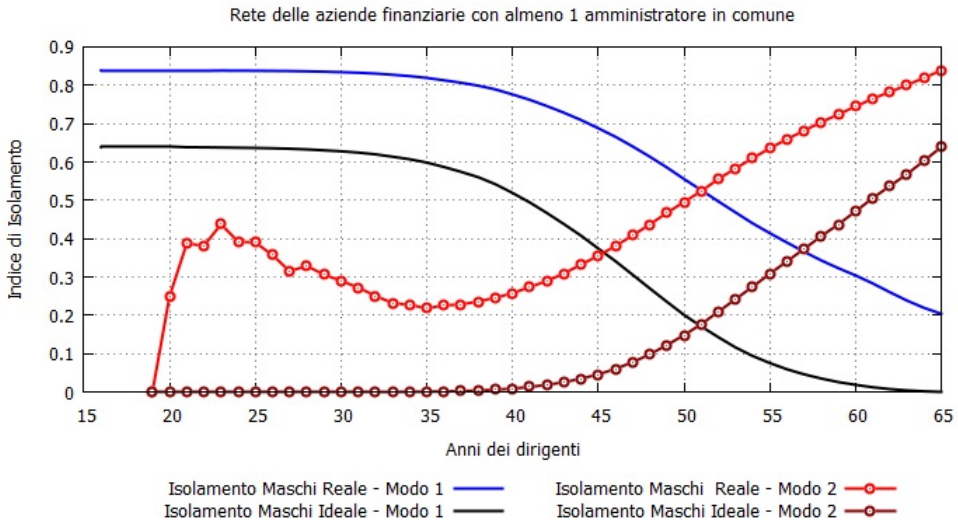


Figura 5.16: Isolamento - Finanza - Maschi - Modo 1 e 2

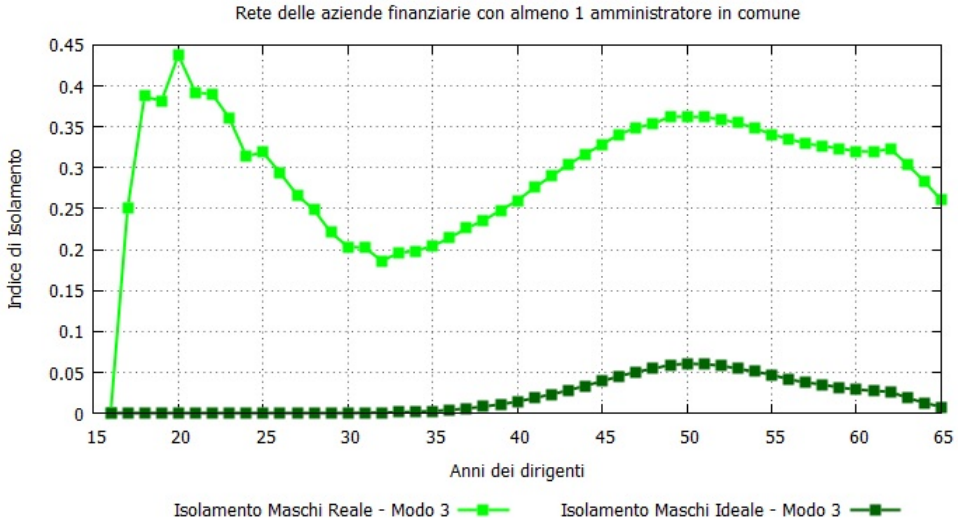


Figura 5.17: Isolamento - Finanza - Maschi - Modo 3

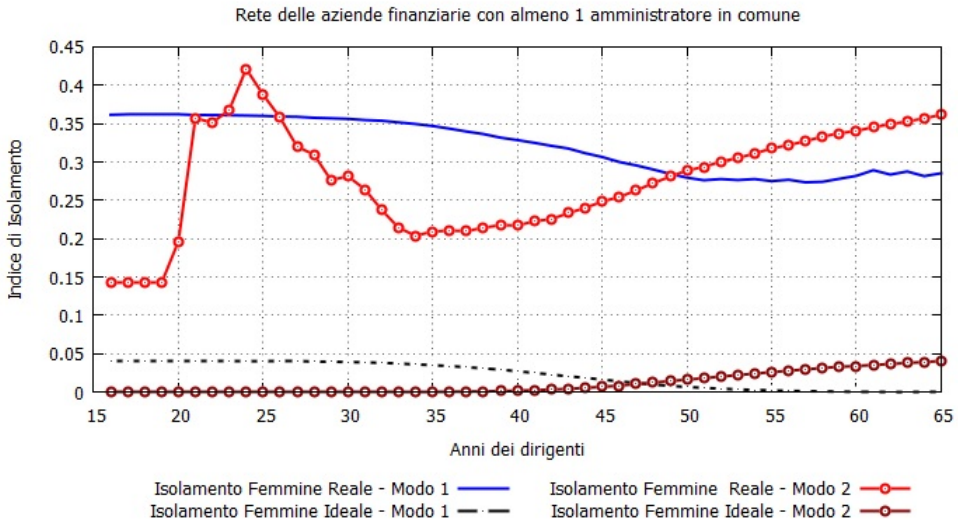


Figura 5.18: Isolamento - Finanza - Femmine - Modo 1 e 2

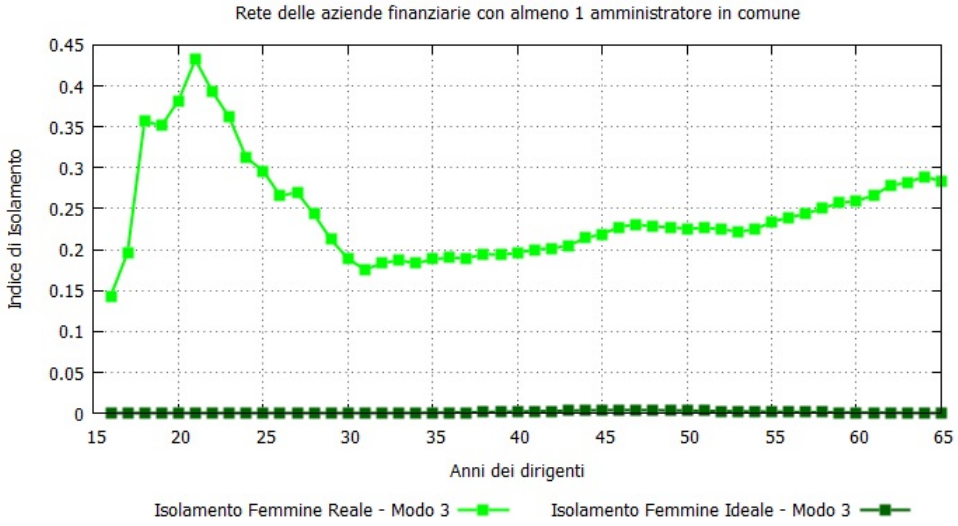


Figura 5.19: Isolamento - Finanza - Femmine - Modo 3

5.2.2 Entropia

Gruppi misti

Nella figura 5.20 sono mostrati i risultati ignorando il sesso.

Le considerazione oggettive che possono essere fatte sono le seguenti:

1. Esiste meno segregazione rispetto al contesto informatico.

$$max(Entropia_{MistiInformatica}) \approx 0.35$$
$$max(Entropia_{MistiFinanza}) \approx 0.7$$

   Nello specifico intervallo [45;55] i livelli di segregazione sono inferiori a quelli di informatica. Un motivo valido può essere perchè i consigli di amministrazione in finanza sono più eterogenei mentre in informatica sembra essere presente il fenomeno di *omofilia*.

Gruppi maschili

Nella figura 5.21sono mostrati i risultati con tutte le modalità di individuazione in cui si considerano gli amministratori di sesso maschili.

Le considerazione oggettive che possono essere fatte sono le seguenti:

1. Vi è meno segregazione rispetto alla rete di informatica.
2. Inoltre possiamo dire che il gruppo appartenente ai [30;40] in proporzione è più equidistribuito rispetto alla rete informatica.

Gruppi femminili

Nella figura 5.22 sono mostrati i risultati con tutte le modalità di individuazione in cui si considerano gli amministratori di sesso femminile.

Le considerazione oggettive che possono essere fatte sono le seguenti:

1. Vi è meno segregazione rispetto alla rete di informatica

$$Entropia_{FemmineInformatica} \approx 0.35$$
$$Entropia_{FemmineFinanza} \approx 0.6$$

2. I maschi sono segregati come le femmine dato da

$$Entropia_{maschi} = Entropia_{femmine} \approx 0.6$$

3. Con l'approccio media mobile si nota come i maschi siano più equidistribuiti rispetto alle femmine. Nonostante in generale le femmine sono presenti nelle unità organizzative come i maschi, le femmine prese per singoli intervalli di età si trovano più isolate rispetto al resto della comunità.

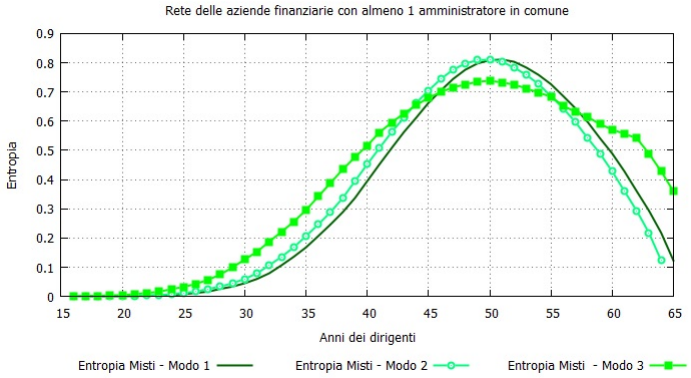


Figura 5.20: Entropia - Finanza - Gruppo Misti

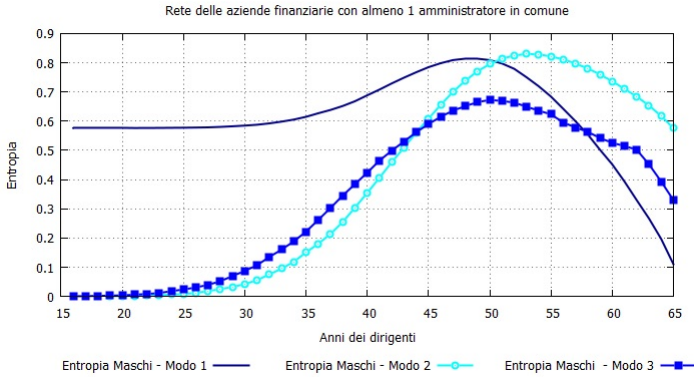


Figura 5.21: Entropia - Finanza - Gruppo Maschili

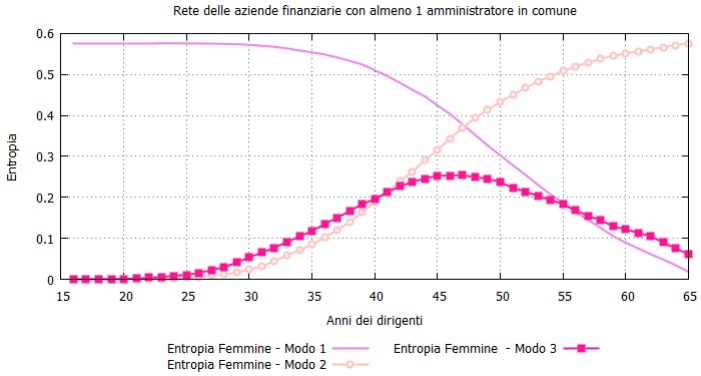


Figura 5.22: Entropia - Finanza - Gruppo Femmine

Conclusioni

Il mondo del lavoro è un tipico ambiente dove si possono sviluppare forme di isolamento, per questo la tesi è nata con scopo di offrire un modo intuitivo di analizzare tale fenomeno all'interno di reti sociali economiche.

Il progetto coinsiste nell'intero processo di data mining per estrarre informazione sulla segregazione in uno specifico tipo di rete, in cui il nodo identifica una azienda italiana e l'arco il numero di amministratori in comune fra due aziende.

Le misure utilizzate sono state l'indice di isolamento e l'entropia nella loro forma binaria. Il sistema trasforma i dati in ingresso in una rete sociale, la scrive su file, individua le comunità ed infine calcola le misure.

Gli indici sono espressi in funzione di un gruppo protetto e calcolati sulle comunità connesse della rete sociale. É stato necessario definire le componenti della rete e l'insieme di gruppi protetti da analizzare.

Infine sono stati commentati i grafici dei risultati, incrociando le informazioni ottenute grazie alle due misure utilizzate. Sono state analizzate due reti, una appartenente al settore informatico, l'altra a quello finanziario.

Viene offerto uno scenario in cui le fasce di amministratori giovani sono più isolate rispetto alle altre fasce, nel settore informatico più che nel finanziario. Le donne di età inferiore ai 30 anni sono più integrate nel settore finanzario. Nel settore informatico le donne hanno più difficoltà ad integrarsi rispetto ai loro colleghi maschi, i quali soprattutto in età giovanile sono isolati rispetto ai loro colleghi di finanza.

In questo breve accenno ai risultati ottenuti si può mostrare come siano emerse alcune forme più o meno accentuate di segregazione, l'onere di interpretarle viene lasciato al lettore o agli analisti del settore.

Il sistema implementato è dotato di una forte espressività della realtà, offrendo due misure di segregazione su 450 gruppi protetti. Inoltre i risultati sono stati progettati affinché sia possibile plottarli in grafici lineari per agevolare l'analisi con un approccio visivo immediato.

Non si nasconde che questo tipo di software si sarebbe prestato ad una tecnica di programmazione parallela conosciuta come *data parallelism* che può consentire miglioramenti nelle performance ma che non è stata implementata perchè i tempi di esecuzione relativamente bassi non lo hanno richiesto.

Una autocritica va mossa per l'utilizzo della libreria *Jung* che occupa la semi totalità del tempo di esecuzione.

Si può affermare che molto può essere ancora fatto in questo ambito di ricerca dallo studio di nuove misure da calcolare all'innesto di nuovi approcci per identificare i gruppi protetti, fino all'estensione dell'approccio per altri tipi di rete sociale.

Appendice A

Manuale d'uso

Manuale d'esecuzione

Il progetto si esegue dalla classe Main.java che utilizza un file di configurazione che deve essere nella solita cartella e che definisce il flusso di esecuzione.

A.1 Funzionalità fornite

Grazie all'utilizzo di un file di configurazione, il sistema offre la possibilità di caricare alcuni parametri che consentono di specificare particolari comportamenti in base alla proprie esigenze:

1. Possibilità di salvare lo stato di esecuzione al momento in cui sono già state calcolate le unità organizzative: per evitare ad ogni esecuzione la creazione del grafo e il calcolo delle comunità. Qualora si scegliesse di attivare questa funzione, va impostato il path del file che contiene le unità organizzative. Il formato di ogni riga di output è il seguente

   $$[\text{UnitàOrganizzativa}_j]$$

   dove UnitàOrganizzativa$_j$ è definito come

   $$[-ID_1 \ldots -ID_n\ ]$$

   in cui esiste un identificatore ID_i $\forall$ azienda i che appartiene all'unità organizzativa j, questa funzione permette di risparmiare quasi il 90% del tempo di esecuzione (By-Passando la libreria Jung).

2. Possibilità di configurare φ ovvero il numero minimo di amministratori che due aziende devono avere in comune per essere considerate collegate nella rete.

3. Possibilità di scegliere l'algoritmo che individua le sottocomponenti di un grafo, fra gli algoritmi "WeakComponent" (l'unico utilizzato, si veda il motivo in sezione 3.6, ed illustrato in sezione 3.6.2) e "Betweenness".

4. Possibilità di impostare dei limiti superiori ed inferiori per l'attributo età (impostazione di un'età massima e minima).

5. Possibilità di impostare a quante cifre di arrotondamento si vogliono arrotondare le misure calcolate.

6. Possibilità di impostare quanti archi con massima "betweenness" rimuovere.

7. Possibilità di impostare per il calcolo del gruppo protetto "Media Mobile - Modo 3" il parametro WS (visto nella sezione 3.4.1).

A.2 Formato dell'output

Nel sistema vi sono varie componenti che producono output su file, di seguito si esprime il formato di tutti i file

GestoreIndiceIsolamento Produce 3 file che rappresentano il valore dell'indice reale ed ideale rispettivamente per i gruppi protetti dei maschi, delle femmine e degli individui misti.

Per cui ogni riga i possiede il formato

$$[età_i \ IsolamentoReale(GP_i) \ IsolamentoIdeale(GP_i)]$$

per cui età$_i$ rappresenta l'età a cui fa riferimento il gruppo protetto GP_i, IsolamentoReale(GP_i) rappresenta il calcolo dell'indice reale di GP_i e $IsolamentoIdeale(GP_i)$ rappresenta il calcolo dell'indice ideale di GP_i.

GestoreEntropia Produce 3 file che rappresentano il valore dell'indice reale per i gruppi protetti dei maschi, delle femmine e degli individui misti. Per cui ogni riga i possiede il formato

$$[età_i \ Entropia(GP_i)]$$

in cui Entropia(GP_i) rappresenta l'entropia per il gruppo protetto GP_i.

GestoreDelGrafo Produce 2 file che contengono gli elementi del grafo costruito

- i nodi (viene data la possibilità di scegliere il formato di output desiderato, l'unico implementato è per il software di analisi *Gephi*) e possiede il seguente formato

$$[ID, Modularity, Weight]$$

in cui *ID* è il codice fiscale identificativo dell'azienda, *Modularity* un valore utile per il software di analisi *Gephi* e *Weight* rappresenta il numero di amministratori dell'azienda.

- gli archi, possiede il seguente formato

$$[Label, Source, Target, Type, Weight]$$

in cui *Label* è l'identificativo dell'arco, *Source e Target* sono gli *ID* delle aziende coinvolte, *Type* è il tipo di arco (sempre "Undirected") e *Weight* è il peso dell'arco, ovvero quanti amministratori le due aziende sourcetarget hanno in comune.

Produce inoltre altri 2 file, uno che misura la quota di ogni gruppo protetto rispetto alla popolazione globale, secondo il tipo di approccio utilizzato (modo 1, 2 o 3) e l'altro che esprime la distribuzione assoluta per ogni età. Il primo scritto nel formato

$$[età_i \ quota(GP_{iFemmine}) \ quota(GP_{iMaschi}) \ quota(GP_{iFemmine}) + quota(GP_{iMaschi}))]$$

per cui età$_i$ rappresenta l'età i a cui fa riferimento il gruppo protetto GP_i, mentre le quote rappresentano le relative quote del gruppo protetto sul totale della popolazione complessiva.
Il secondo scritto nel formato

$$[età_i \ Maschi(età_i) \ Femmine(età_i) \ Maschi(età_i)+Femmine(età_i)]$$

per cui vengono mostrare le distribuzioni degli individui in base all'età.

Bibliografia

[1] Eurobarometro 349. Attitudes on data protection and electronic identity in the european union. Jun 2011.

[2] Marc Baudoin. *Impara LaTeX*. ´Ecole Nationale Sup´rieure de Techniques Avanc´es.

[3] Stefano Boni and direzione scientifica di Marcello Flores. *Diritti umani : cultura dei diritti e dignità della persona nell'epoca della globalizzazione*. 2007.

[4] Antoni Calvo-Armengol and Matthew O. Jackson. The effects of social networks on employment and inequality. *The American Economic Review*, 94(3):426–454, 2004.

[5] W. A. V. Clark. Residential Preferences and Neighborhood Racial Segregation: A Test of the Schelling Segregation Model. *Demography*, 28(1):1–19, 1991.

[6] G. M. Duncan and D. E Duncan. A methodological analysis of segregation. *American Sociological Review*, 20:210–217, 1955.

[7] Henry G. Duranton, Gilles & Overman. Testing for Localization Using Micro-Geographic Data. *CEPR Discussion Papers*, (3379), 2002.

[8] Federico Echenique and Roland G. Fryer. A Measure of Segregation Based on Social Interactions. *The Quarterly Journal of Economics*, 122(2):441–485, 2007.

[9] Comunità Europea. Carta dei diritti fondamentali dell'Unione Europea. *Gazzetta ufficiale della Comunità Europea*, page C 364/13, 2000.

[10] Mark Fossett. Ethnic Preferences, Social Distance Dynamics, and Residential Segregation: Theoretical Explorations Using Simulation Analysis*. *The Journal of Mathematical Sociology*, 30(3-4):185–273, 2006.

[11] Linton C. Freeman. Segregation in social network. *Sociological Methods & Research*, 6(4):411–430, 1978.

[12] Mark Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(1360):80, 1973.

[13] D. R. James and K. E. Tauber. Measures of segregation. *Sociological Methodology*, 13:1–32, 1985.

[14] Young Chul Kim. Lifetime network externality and the dynamics of group inequality. Technical Report 18767, University Library of Munich, Germany, 2009. http://ideas.repec.org.

[15] U.S. Federal Legislation. Equal Credit Opportunity Act, 2012.

[16] Angelo Mele. A Structural Model of Segregation in Social Networks. *NET Institute Working*, 10-16(3379), Sept 2010.

[17] Lynn Smith-Lovin Miller McPherson and James M Cook. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27, Aug 2001.

[18] James D Montgomery. Social networks and labor-market outcomes: Toward an economic analysis. *American Economic Review*, 81(5):1407–1418, 1991.

[19] M. E. J. Newman. Assortative Mixing in Networks. *Phys. Rev. Lett.*, 89:208701, Oct 2002.

[20] M. E. J. Newman. Mixing patterns in networks. *Phys. Rev. E*, 67:026126, Feb 2003.

[21] Romans Pancs and Nicolaas J. Vriend. Schelling's spatial proximity model of segregation revisited. *Journal of Public Economics*, 91(1-2):1–24, 2007.

[22] Anthony H. Pascal. *The Economics of Housing Segregation.* 1967.

[23] Michael J. Piore Peter B. Doeringer. *Internal Labor Markets and Manpower Analysis.* 1971.

[24] Albert Ress. Information networks in labor markets. *American Economic Review*, 1966.

[25] Tim Rogers and Alan J. McKane. A unified framework for Schelling's model of segregation. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(7):P07006, 2011.

[26] Salvatore Ruggieri and Andrea Romei. A multidisciplinary survey on discrimination data analysis. Technical report, Dipartimento di Informatica, Largo B. Pontecorvo 3, 56127 Pisa, Italy.

[27] T C Schelling. Dynamic models of segregation. *Journal of Mathematical Sociology*, 1(May):143–186, 1971.

[28] Marco van der Leij, Meredith Rolfe, and Ott Toomet. Social networks and the economic performance of minorities, 2009. Working Paper, http://equalsoc.org.