Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea in Matematica

L'omologia persistente applicata allo studio delle reti complesse

Tesi di Laurea in Topologia Algebrica

Relatore:
Prof. Massimo Ferri

Presentata da:
Matteo Giusti

Sessione Unica
Anno Accademico 2017-2018

Scusate il ritardo...

Indice

0.1 Introduzione

In un mondo sempre più interconnesso, formato da sistemi complessi collegati tra loro nei modi più disparati, abbiamo a disposizione una immensa mole di dati in continuo aggiornamento che cresce ad una velocità senza precedenti. Fortunatamente lo sviluppo di nuovi metodi sperimentali e di nuove tecnologie computazionali estremamente potenti va di pari passo con la produzione di dati. Siamo così in grado di studiarli e analizzarli per riuscire ad estrapolare conoscenze ed informazioni, distinguendo quelle utili da quelle infruttuose: è l'era dei big data.
Se prima gli analisti lavoravano con dati ordinati, regolari o preimpostati, ora sono in grado di trattare con dataset notevolmente eterogenei grazie agli incredibili progressi fatti sul campo; pensiamo agli algoritmi di autoapprendimento (machine learning) oppure ai computer quantici, IBM ne ha recentemente commercializzato un primo modello.
In questo elaborato si discute quindi di come uno strumento straordinariamente versatile stia apportando notevoli contributi alla risoluzione di problemi di classificazione e recupero dei dati: l'omologia persistente.
Questa branca della topologia algebrica nasce negli anni '90 con il nome di teoria di taglia dal lavoro di Patrizio Frosini e Massimo Ferri, professori dell'università di Bologna, e si sviluppa in collaborazione con il gruppo di A. Verri a Genova. Indipendentemente, ma successivamente, le stesse idee fioriscono anche alla Boulder University, Colorado, con il lavoro di V. Robins. Per maggiori informazioni si vedano gli articoli [5], [12].
La ricerca in questo campo ha portato negli ultimi anni ad importanti successi; è notevole soprattutto come l'applicazione dell'omologia persistente spazi tra gli argomenti più disparati: leucociti, nei e melanomi, cicloni, galassie, firme, alfabeto dei segni, connessioni celebrali, profili umani etc. Nell'articolo [3] se ne danno interessanti esempi.
Nel primo capitolo si introducono quindi le nozioni matematiche necessarie per un primo approccio con l'omologia persistente. Si parlerà anche di grafi e simplessi, concetti utili per la modellizzazione dei dati reali.
Nel secondo capitolo si presentano le reti complesse, ottime strutture per la sintesi e la strumentalizzazione dei dati. Segue un'analisi di queste ultime attraverso l'omologia persistente che risulterà essere un ragionevole strumento discriminante, anche quando le reti sembrano ad un primo sguardo notevolmente simili. Oltre queste considerazioni il terzo capitolo presenta anche un paragone con le proprietà topologiche non persistenti.
L'ultimo capitolo si discosta invece dall'omologia per parlare brevemente di una proprietà intrinseca delle reti sociali, l'assortatività del secondo ordine.

Capitolo 1

Premesse matematiche

Se si vuole applicare la geometria all'analisi delle forme e ai problemi di classificazione e recupero dei dati spesso basta associare ad ogni immagine i descrittori di forma, stringhe di misure geometriche, per poi applicare, su di essi piuttosto che sull'immagine, una trasformazione, che sia euclidea, affine o proiettiva. Questo strumento dell'algebra matriciale funziona bene sullo studio di oggetti rigidi ma risulta inconveniente se si prendono in considerazione immagini di origine naturale. Per esempio non c'è una trasformazione geometrica di questo tipo che porti un uomo dalla posizione eretta all'essere seduto. Si cambia allora strumento: gli omeomorfismi, funzioni biettive e bicontinue tra due spazi topologici. Se si vuole quindi associare un uomo seduto ad uno in piedi gli omeomorfismi risultano lo strumento adatto. La loro particolare duttilità risulta però un problema qualora volessimo distinguere in che posa si trova l'uomo. In entrambi i casi il primo problema da affrontare è sapere se due spazi sono omeomorfi. Si adopera allora la topologia algebrica che associa ad ogni spazio i rispettivi invarianti topologici, proprietà che si conservano tra spazi omeomorfi. Un esempio sono i numeri di Betti: $\beta_0(X)$ ci dice il numero delle componenti connesse per archi, $\beta_1(X)$ il numero di buchi (come quello di una ciambella), $\beta_2(X)$ conta i vuoti bidimensionali (come la camera d'aria di un pallone) e così via. Riprendendo l'esempio precedente per un topologo l'uomo seduto e quello in piedi sono la stessa cosa. Per poterli distinguere utilizziamo allora la topologia persistente: associamo ad uno spazio topologico X una funzione continua, detta funzione filtrante, che esprime il concetto di sagoma o il punto di vista dell'osservatore. Risolviamo così il problema della posa: cercando i descrittori di forma negli insiemi di sottolivello riusciamo infine a distinguere l'uomo seduto da quello in piedi.
Passiamo quindi ad enunciare sia le nozioni ora presentate sia quelle che utilizzeremo nel corso dell'elaborato. Quella che segue è una descrizione intuitiva dei concetti e non può di certo sostituire un opportuno manuale se si vuole comprendere a fondo gli argomenti. Per la stesura di questo capitolo si è fatto riferimento agli articoli [3], [4], [1].

1.1 Grafi e complessi simpliciali

1.1.1 Teoria dei grafi

Un *grafo* è un diagramma che permette di schematizzare una grande varietà di situazioni e processi. È formato da elementi detti *nodi o vertici* che possono essere connessi tra loro attraverso collegamenti chiamati *archi*, *lati o spigoli*.
Formalmente si dice grafo una coppia ordinata $G = (V, E)$ alla quale è associata una funzione ψ.
V rappresenta l'insieme dei nodi ed E è l'insieme degli archi, disgiunto da E. ψ si chiama funzione di incidenza ed associa ad ogni arco, elemento di E, una coppia non ordinata di vertici, elementi di V, non necessariamente distinti.
Se b è un arco e x e y sono vertici tali che $\psi(b) = \{x, y\}$ allora si dice che b collega x e y e che questi vertici sono gli estremi dell'arco.
Un arco che ha i due estremi coincidenti si dice cappio o self-loop.
Se la funzione di incidenza associa ad ogni arco una coppia ordinata di vertici il grafo si dice orientato.

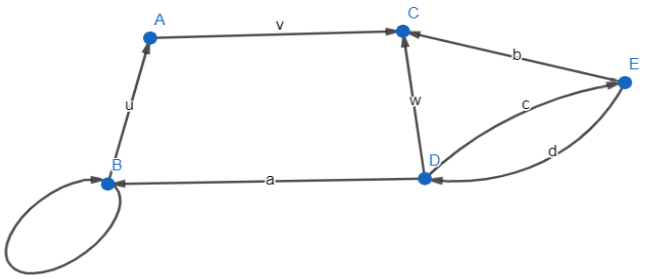


Figura 1.1: Esempio di grafo orientato con un cappio e un lato multiplo.

Un grafo si dice *completo* se due qualsiasi dei suoi vertici sono connessi da un arco.
Se vengono associati ad ogni nodo ed arco del grafo dei pesi o valori, allora il grafo si dice *pesato*.

Un grafo non orientato si dice *semplice* se non contiene cappi né lati multipli.

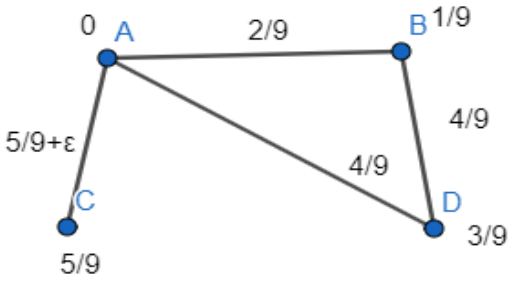


Figura 1.2: Esempio di grafo semplice pesato.

La *matrice di adiacenza* è una particolare struttura dati comunemente utilizzata nella rappresentazione dei grafi. Se G è un grafo con $V = \{v_1, v_2, \ldots, v_n\}$ la matrice di adiacenza A è così definita:

$$a_{ij} = \begin{cases} 1 & \textit{se esiste un arco che connette } v_i \textit{ con } v_j \\ 0 & \textit{altrimenti} \end{cases}$$

Se il grafo non è orientato la matrice di adiacenza A è simmetrica.
Un *ipergrafo* è un grafo in cui un arco può essere collegato a un qualunque numero di vertici.
Nella Figura 1.3 l'arco che collega i nodi A,B,D è rappresentato dall'area in grigio.

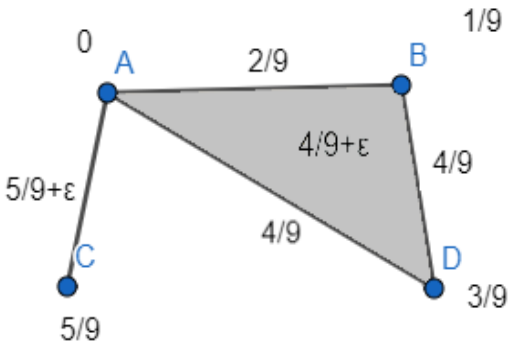


Figura 1.3: Un ipergrafo

1.1.2 Simplessi

Definizione 1.1 (**Simplesso**)**.**
Si definisce un p-simplesso σ come l'inviluppo convesso, in uno spazio euclideo, di un insieme di $p+1$ punti, detti vertici, che non sono contenuti in un sottospazio euclideo (p - 1)-dimensionale. Diciamo che il simplesso è generato dai suoi vertici e chiameremo faccia di un simplesso σ il simplesso τ generato da un sottoinsieme non vuoto di vertici di σ. In tal caso scriviamo $\tau \leq \sigma$.

Definizione 1.2 (**Complesso simpliciale**)**.**
Un complesso simpliciale è una collezione finita K di simplessi di un dato spazio Euclideo tali che:
1. Se $\sigma \in K$ e $\tau \leq \sigma, \Rightarrow \tau \in K$.
2. Se $\sigma_1, \sigma_2 \in K$ e $\sigma_1 \cap \sigma_2 \neq \emptyset \Rightarrow \sigma_1 \cap \sigma_2$ è faccia sia di σ_1 sia di σ_2.
Lo spazio di un complesso simpliciale K è il sottospazio topologico dello spazio euclideo $|K|$ costituito dall'unione dei tutti i simplessi di K.

Definizione 1.3 (p**-bordi e** p**-cicli**)**.**
Dato un complesso simpliciale finito K, si chiama p-catena ogni combinazione lineare formale di p-simplessi con coefficienti in $\mathbb{Z}_2$:

$$\sum_i a_i \sigma_i$$

dove $a_i \in \mathbb{Z}_2$ e σ_i è un p-simplesso. Le p-catene formano lo spazio vettoriale C_p a coefficienti in $\mathbb{Z}_2$.
Ogni p-catena identifica un insieme di p-simplessi di K e la somma di due p-catene corrisponde alla differenza simmetrica dei rispettivi insiemi.
L'operatore di bordo è una trasformazione lineare $\partial_p : C_p \to C_{p-1}$, per ogni $p \in \mathbb{Z}$. È sufficiente definire tale operatore sui p-simplessi e poi estenderlo per linearità alle p-catene. Scriviamo il p-simplesso σ come $\sigma = [u_0, u_1, \ldots, u_p]$ e indichiamo con $[u_0, \ldots, u'_j, \ldots, u_p]$ la faccia di σ generata da tutti i suoi vertici eccetto u_j, per $j \in \{0, \ldots, p\}$. Definiamo poi

$$\partial_p(\sigma) = \sum_{j=0}^{p} [u_o, \ldots, u'_j, \ldots, u_p]$$

È possibile dimostrare che $\partial_p \partial_{p+1} = 0$, cosicché $B_p = Im\partial_{p+1}$ sarà contenuto in $Z_p = Ker\partial_p$. Gli elementi di B_p sono chiamati p-bordi e gli elementi di Z_p si dicono p-cicli.

Definizione 1.4 (Gruppo di omologia).
Il p-esimo gruppo di omologia di un simplesso K è definito come il quoziente

$$H_p(K) = Z_p(K)/B_p(K)$$

Le classi di omologia sono rappresentate come cicli che non sono dei bordi. Due cicli sono omologhi se la loro differenza è un bordo.

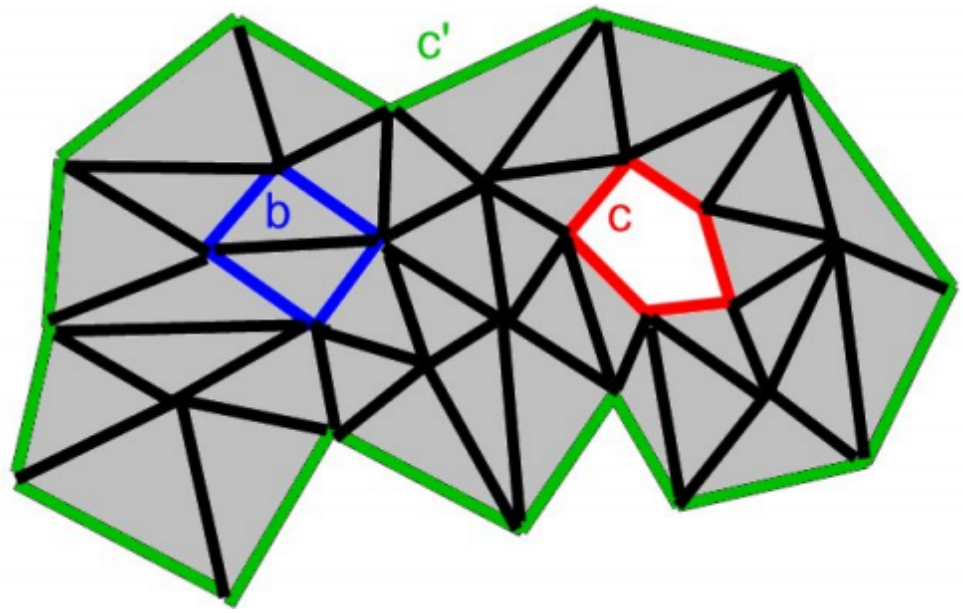


Figura 1.4: Cicli e bordi

Vediamo un esempio di quanto appena detto nella Figura 1.4: vi è rappresentato un complesso simpliciale K composto dai triangoli in grigio e dalle loro facce. In blu è segnata la catena b che è un 1-ciclo ma anche un bordo; la catena rossa c e quella verde c' sono 1-cicli ma non sono bordi. c e c' sono omologhe.

1.2 Omologia persistente

Ricordiamo il concetto di omeomorfismo con il classico esempio della tazza e della ciambella.

Definizione 1.5 (**Omeomorfismo**).
Dati due spazi topologici X e Y un omeomorfismo da X a Y è una funzione continua, invertibile, con inversa continua. Se esiste un omeomorfismo tra due spazi topologici questi si dicono omeomorfi.

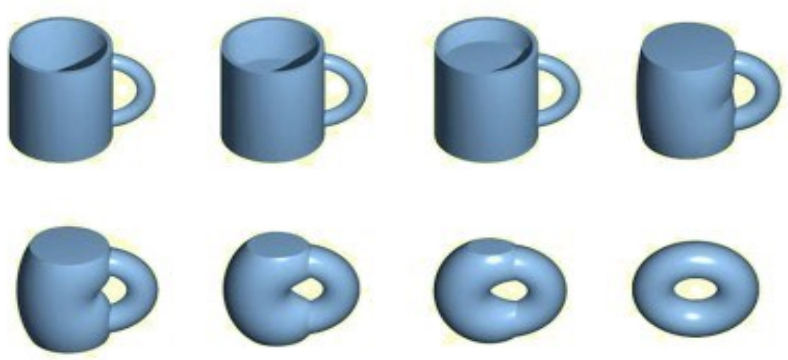


Figura 1.5: Trasformazione di una tazza in una ciambella

Nella Figura 1.5 si può vedere come una tazza è omeomorfa ad una ciambella, esiste cioè una deformazione continua che trasforma un oggetto nell'altro e viceversa.

Definizione 1.6 (**Numeri di Betti**).
Per $k \in \mathbb{N} \cup \{0\}$, il k-esimo numero di Betti $\beta_k(X)$ è la dimensione del k-esimo gruppo di omologia $H_k(X)$, cioè il numero di generatori indipendenti (le classi di omologia dei k-cicli) di questo spazio.

Definizione 1.7 (**Funzione filtrante o funzione misurante**).
Sia X uno spazio topologico e sia $f : X \to \mathbb{R}$ una funzione continua.
(X, f) si dice coppia di taglia e la funzione f funzione filtrante.

Definizione 1.8 (**Insieme di sottolivello**).
Sia (X, f) una coppia di taglia e sia $u \in \mathbb{R}$
l'insieme $X_u = \{x \in X | f(x) \leq u\}$ si chiama insieme di sottolivello.

Riprendendo l'esempio della tazza e della ciambella una possibile funzione filtrante potrebbe essere l'altezza. Nella Figura 1.6 sono rappresentati gli insiemi di sottolivello alle altezze a',b',c.

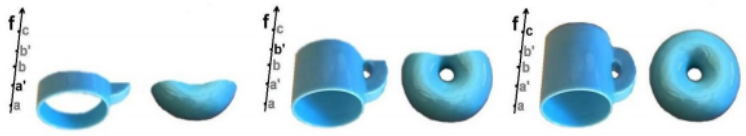


Figura 1.6: Tre insiemi di sottolivello di una tazza e di una ciambella

Definizione 1.9 (**k-esimo gruppo di omologia persistente**).
Per ogni $u, v \in \mathbb{R}$ con $u < v$ la mappa di inclusione $\iota_{(u,v)} : X_u \to X_v$ è continua ed induce, qualunque sia il grado k, una trasformazione lineare

$$\iota^{*,k}_{(u,v)} : H_k(X_u) \to H_k(X_v)$$

Si definisce il k-esimo gruppo di omologia persistente con

$$PH_k(u, v) = \iota^{*,k}_{(u,v)}(H_k(X_u)) \subseteq H_k(X_v)$$

Definizione 1.10 (**Numeri di Betti persistenti**).
Si costruisce la funzione k-PBN assegnando ad ogni coppia (u, v) il numero $dim(Im(\iota^{*,k}_{(u,v)}))$ che si chiama k-esimo numero di Betti persistente.
La funzione k-PBN conta il numero di classi di k-cicli di $H_k(X_u)$ che sopravvivono, persistono, in $H_k(X_v)$

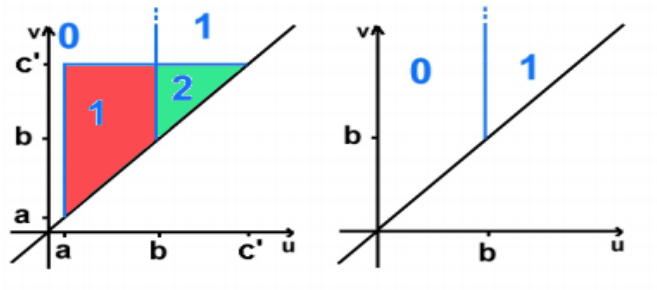


Figura 1.7: 1-PBN della tazza a sinistra e della ciambella a destra

Ricollegandoci all'esempio della Figura 1.6 calcoliamo i numeri di Betti persistenti di grado 1, riportati nella Figura 1.7. Per quanto riguarda la tazza messa sottosopra (a sinistra) al livello di nascita a è già presente un 1-ciclo, dovuto alla cavità dell'oggetto, che scomparirà all'altezza della formazione del fondo, cioè il livello c'. Al livello b si forma un altro 1-ciclo, dovuto al buco del manico, che non scomparirà più.
Osservando il grafico l'area in rosso ci dice che l'1-ciclo dei sottolivelli $X_{a'}$ con $a \leq a' < b$ persiste nei sottolivelli $X_{b'}$ con $a' < b' < c'$ per poi morire in $X_{c'}$.
L'area in verde ci dice che i due 1-cicli di $X_{b'}$ con $b \leq b' < c'$ persistono in ogni sottolivello $X_{b''}$ con $b' \leq b'' < c'$. Oltre il livello c' si conserva solo l'1-ciclo del manico.
La funzione 1-PBN della ciambella (a destra) ci dice che al livello b nasce l'unico 1-ciclo dell'oggetto che persiste all'infinito.

Definizione 1.11 (**Diagrammi di persistenza**)**.**
Le funzioni k-PBN sono interamente determinate dalla posizione di alcuni punti e linee di discontinuità, cioè i cornerpoints e le cornerlines (o cornerpoints all'infinito). Un cornerpoint di coordinate (u, v) rappresenta il livello di nascita u e il livello di morte v di un ciclo. L'ascissa di una cornerline rappresenta il livello di nascita di un ciclo che non muore mai.
Si chiama persistenza di un cornerpoint la differenza $v - u$ delle sue coordinate.
Questi punti e queste linee formano il k-diagramma di persistenza, definito con k-PD.

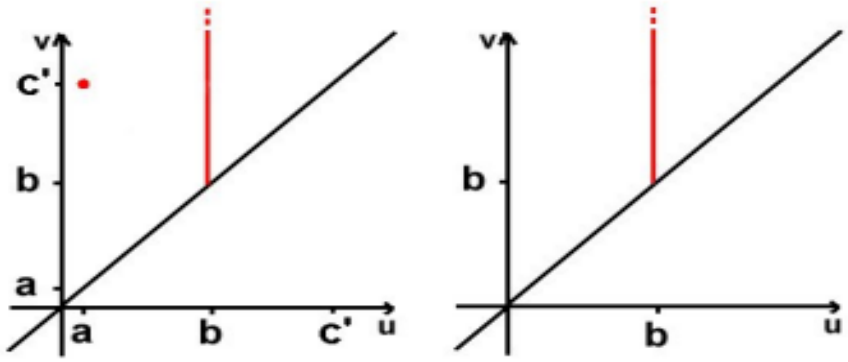


Figura 1.8: 1-PD della tazza a sinistra e della ciambella a destra

Definizione 1.12 (**Distanza di matching**).
Siano (X, f) e (Y, g) due coppie di taglia e siano $\mathcal{D}_{X,f}$ e $\mathcal{D}_{Y,g}$ i rispettivi k-PD. Per confrontare i due diagrammi di persistenza si comincia definendo il costo diretto per comparare due qualsiasi elementi $q = (q_n, q_m) \in \mathcal{D}_{X,f}$ e $q\prime = (q'_n, q'_m) \in \mathcal{D}_{Y,g}$ come la norma-∞ della loro differenza

$$\|q - q\prime\|_\infty := max(|q_n - q'_n|, |q_m - q'_m|)$$

dove il pedice n sta per nascita e m per morte.
Osserviamo che l'ordinata di una cornerline è uguale a ∞. Si risolve questo problema normalizzando i valori, cioè si va a considerare un ciclo che non muore mai come un cornerpoint di ordinata uguale a 1. Osserviamo poi che la diagonale di un diagramma di persistenza può essere considerata come l'insieme di infiniti cornerpoints con tempo di nascita uguale al tempo di morte. Si può quindi comparare un punto q di un k-PD con la propria proiezione q_p sulla diagonale. In questo caso risulta $\|q - q_p\|_\infty = \frac{|q_m - q_n|}{2}$.
Si definisce allora il costo di accoppiamento $c(q, q\prime)$ come il minimo tra il costo diretto e il massimo tra $\frac{|q_m - q_n|}{2}$ e $\frac{|q'_m - q'_n|}{2}$.
La distanza di matching, o bottleneck distance, dei due diagrammi $\mathcal{D}_{X,f}$ e $\mathcal{D}_{Y,g}$ è l'inf dei costi di accoppiamento calcolato considerando tutti i possibili matching.
Indichiamo con b_k la distanza di matching tra $k - PDs$.

Definizione 1.13 (**Pseudodistanza naturale**).
Date due coppie di taglia (X, f) e (Y, g) con X e Y omeomorfi, il costo di un dato omeomorfismo $\theta : X \to Y$ è il

$$\sup_{x \in X} |g(\theta(x)) - f(x)|$$

.
La pseudodistanza naturale di (X, f) e (Y, g) è l'inf di questi costi calcolato tra tutti i possibili omeomorfismi.

Teorema 1.2.1. *Date due coppie di taglia (X, f) e (Y, g) e i rispettivi diagrammi di persistenza k-PD, si ha che la distanza di matching tra $\mathcal{D}_{X,f}$ e $\mathcal{D}_{Y,g}$ è un limite inferiore della pseudodistanza naturale tra (X, f) e (Y, g).*
Oltretutto risulta essere la migliore approssimazione che possiamo ottenere.

Capitolo 2

Networks

2.1 High-order Networks

Possiamo pensare a un generico dataset con i rispettivi collegamenti interni che lo caratterizzano come un high-order network, cioè un ipergrafo pesato.
A sua volta possiamo considerare questo ipergrafo come un complesso simpliciale.
Applicata per comparare reti con un basso numero di nodi, la pseudodistanza naturale ha avuto successo come strumento discriminante. Questa distanza è calcolata considerando tutte le possibili corrispondenze tra i nodi, al crescere di questi ultimi risulta allora computazionalmente inaffrontabile.
Questo problema si risolve studiando il parallelismo tra high-order networks e le filtrazioni, con relative caratteristiche omologiche, della topologia algebrica: si ha che la distanza di matching, facilmente elaborabile da un computer, risulta essere un limite inferiore della distanza tra network. Si è ottenuto così una sua buona approssimazione, considerando che i limiti superiori sono facili da determinare.
Questo capitolo si basa sugli studi di H. Huang pubblicati in [7], [6].

Definizione 2.1 (**Network**).
Un network N_X^K di ordine K definito sull'insieme dei nodi X è una collezione di $K+1$ funzioni di correlazione normalizzate

$$\{r_X^k : X^{k+1} \longrightarrow [0,1]\}_{k=0}^K$$

dove X^{k+1} è lo spazio dei vettori $k+1$ dimensionali le cui componenti sono elementi di X
e $[0,1]$ è l'intervallo unitario chiuso.

Per ogni elemento $x_{0:k} := (x_0, x_1, ..., x_k)$ di X^{k+1} il valore $r_X^k(x_{0:k})$ si può intendere come una misura di similarità o dissimilarità tra gli elementi del gruppo.

Si andranno a considerare solo i networks che rispettano determinate proprietà di simmetria ed identità cosicché le relazioni r_X^k non dipendano dall'ordine e dalle ripetizioni degli elementi dei vettori. Cioè, per esempio, si vuole avere $r_X^2(w,w,y) = r_X^1(w,y)$ e $r_X^2(w,y,z) = r_X^2(y,z,w)$.
Formalmente parlando:
Per la proprietà di simmetria deve valere

$$r_X^k(x_{0:k}) = r_X^k(x_{[0:k]})$$

dove $x_{[0:k]}$ è un arbitrario riordinamento di $x_{0:k}$.
Per la proprietà di identità deve valere

$$r_X^k(x_{0:k}) = r_X^{k'}(x_{l_0:l_{k'}})$$

$\forall$ vettore $x_{0:k}$ e $\forall$ ordine k,
dove $x_{l_0:l_{k'}}$ è un sottovettore tale che il numero di elementi unici è lo stesso di $x_{0:k}$.
Si denota con $\mathcal{N}_X^K$ l'insieme di tutti gli high-order networks.

Definizione 2.2 (**Network k-isomorfi**)**.**
Si dice che due networks N_X^K e N_Y^K sono k-isomorfi se esiste una biiezione
$\pi : X \longrightarrow Y$ tale che $\forall\, x_{0:k} \in X^{k+1}$ si ha

$$r_Y^k(\pi(x_{0:k})) = r_X^k(x_{0:k}) \tag{2.1}$$

Condizione necessaria affinché la (2.1) sia soddisfatta è che X ed Y siano equipotenti.
Si indica che i networks N_X^K e N_Y^K sono k-isomorfi con la notazione $N_X^K \cong_k N_Y^K$

Quozientando lo spazio $\mathcal{N}_X^K$ con questa relazione di equivalenza otteniamo $\mathcal{N}^K mod_{\cong_k}$, lo spazio dei K-order networks in cui i network k-isomorfi sono rappresentati dallo stesso elemento.
Per ogni $0 \leq k \leq K$ lo spazio $\mathcal{N}^K mod_{\cong_k}$ può essere dotato di una pseudometrica.
Per definire questa "distanza" dobbiamo prima introdurre il concetto di corrispondenza.

Definizione 2.3 (**Corrispondenza**)**.**
Una corrispondenza tra due insiemi X e Y è un sottoinsieme $C \subseteq X \times Y$ tale che
$\forall\ x \in X\ \exists\ y \in Y\ t.c.\ (x,y) \in C$
e $\forall\ y \in Y\ \exists\ x \in X\ t.c\ (x,y) \in C$.
L'insieme di tutte le corrispondenze è $\mathcal{C}(X,Y)$.

Definizione 2.4 (**La pseudometrica $\Gamma_{X,Y}^k(C)$ **)**.**
Dati N_X^K e N_Y^K, una corrispondenza $C \in \mathcal{C}(X,Y)$ e un intero $0 \leq k \leq K$
definisco in $\mathcal{N}^K mod_{\cong_k}$ la seguente pseudometrica

$$\Gamma_{X,Y}^k(C) := \max_{(x_{0:k},y_{0:k})\in C} \mid r_X^k(x_{0:k}) - r_Y^k(y_{0:k}) \mid \tag{2.2}$$

Definizione 2.5 (**Distanza tra Networks**).
Possiamo ora definire la distanza tra due k-order networks N_X^k e N_Y^k:

$$d_{\mathcal{N}}^k(N_X^k, N_Y^k) := \min_{C \in \mathcal{C}(X;Y)} \{\Gamma_{X,Y}^k(C)\} \tag{2.3}$$

Notiamo che $d_{\mathcal{N}}^k(N_X^k, N_Y^k)$ è stata definita senza restrizioni sulla cardinalità dei due insiemi dei nodi X e Y.
La pseudodistanza in (2.2) è stata definita per un fissato intero k. Si vanno allora a generalizzare i concetti precedenti svincolandosi da questa ipotesi.

Definizione 2.6 (**Networks isomorfi**).
Due K-order networks N_X^K e N_Y^K si dicono isomorfi se esiste una biiezione $\pi : X \longrightarrow Y$ tale che la (2.1) valga $\forall\, k$ con $0 \leq k \leq K$ e $\forall\, x_{0:k} \in X^{k+1}$.
Quando due networks N_X^K e N_Y^K sono isomorfi si scrive $N_X^K \cong N_Y^K$
e $\mathcal{N}^K mod_{\cong}$ è il relativo spazio quoziente.

Definizione 2.7 (**Differenza tra Networks**).
Siano dati: due networks N_X^K e N_Y^K, una corrispondenza C su gli insiemi dei nodi X e Y, una p-norma vettoriale $\|\cdot\|_p$. Si definisce così la differenza tra network relativa a C:

$$\|\Gamma_{X,Y}^K(C)\|_p := \|(\Gamma_{X,Y}^0(C), \Gamma_{X,Y}^1(C), \ldots, \Gamma_{X,Y}^K(C))^t\|_p \tag{2.4}$$

nella quale $\Gamma_{X,Y}^k(C)$ è la pseudodistanza di ordine k definita in (3.2).

Definizione 2.8 (p-**Distanza tra Networks**).
La p-distanza di network tra N_X^K e N_Y^K è definita come

$$d_{\mathcal{N},p}(N_X^K, N_Y^K) := \min_{C \in \mathcal{C}(X,Y)} \{\|\Gamma_{X,Y}^K(C)\|_p\} \tag{2.5}$$

2.2 Dissimilarity networks

La funzione di correlazione di grado k, r_X^k, di un network N_X^K non impone restrizioni sulle altre funzioni r_X^l di diverso ordine l dello stesso network N_X^K. D'altro canto è più pratico andare ad utilizzare correlazioni che forniscono un valore più alto o più basso andando ad aumentare il numero di nodi del vettore. Il primo caso che andiamo ad analizzare è quello in cui il valore $r_X^k(x_{0:k})$ si traduce come un livello di dissimilarità tra i nodi.

Definizione 2.9 (**Dissimilarity Networks**).
Si dice che il K-order network $D_X^K = (X, r_X^0, \ldots, r_X^K)$ è un dissimilarity network se per un qualsiasi grado $1 \leq k \leq K$ e per ogni vettore $x_{0:k} \in X^{k+1}$ si ha

$$r_X^k(x_{0:k}) \geq r_X^{k-1}(x_{0:k-1}) \tag{2.6}$$

e l'uguaglianza vale se e solo se tutti gli elementi di $x_{0:k}$ compaiono in $x_{0:k-1}$.
Si denota l'insieme di tutti i dissimilarity networks di ordine K con $\mathcal{D}^K$.

Si considera ora questo esempio: data una comunità di ricerca scientifica si analizzano le dinamiche nel tempo della sua formazione.
Le funzioni di correlazione segnano in questo network l'istante di tempo normalizzato in cui i membri di un dato vettore scrivono il loro primo saggio insieme.
Per $k = 0$ le difformità r^0 misurano il momento in cui gli autori pubblicano il loro primo saggio senza cooperare con gli altri. Nella Figura 2.1 gli autori A,B,C,D pubblicano per la prima volta rispettivamente al tempo 0 , 1/9 , 5/9 , 3/9.

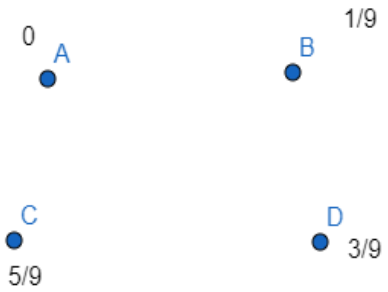


Figura 2.1: I valori r^0_X del gruppo di coautori

Il valore r^1_X degli archi tra le coppie indica il momento in cui i nodi diventano co-autori. Si noti che vale sempre $r^1(x, y) \geq r^0(x)$ e $r^1(x, y) \geq r^0(y)$. Nella Figura 2.2 si vede che A e B scrivono il primo articolo insieme al tempo 2/9, maggiore sia di 0 sia di 1/9.

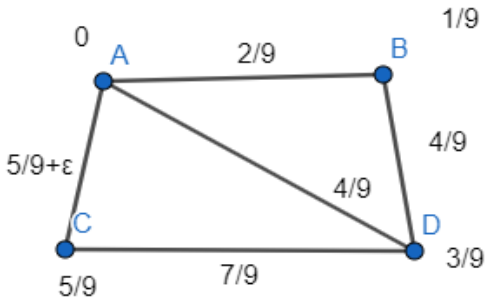


Figura 2.2: I valori r^1_X del gruppo di coautori

I due valori di dissimilarità r_X^2 indicano il momento in cui le triplette A,C,D e A,B,D diventano coautori. Anche in questo caso vale $r_X^2(w, y, z) \geq r_X^1(w, y)$.

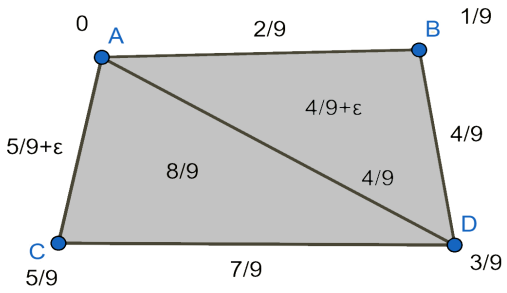


Figura 2.3: I valori r_X^2 del gruppo di coautori

Una costante ε è aggiunta per rendere $r_X^1(A; C)$ strettamente maggiore sia di $r_X^0(A)$ sia di $r_X^0(C)$, e per rendere $r_X^2(A, B, D)$ strettamente maggiore sia di $r_X^1(A, D)$ sia di $r_X^1(B, D)$. È un'operazione obbligatoria perché l'uguaglianza in (2.6) vale se e solo se gli elementi di $x_{0:k}$ sono tutti e soli gli elementi di $x_{0:k-1}$.

2.2.1 Passaggio all'omologia persistente

Passiamo dalla teoria dei grafi alla topologia algebrica, si va cioè a rappresentare un high-order network come un complesso simpliciale. Ad ogni simplesso è associato il relativo peso dato dalle dissimilarità.

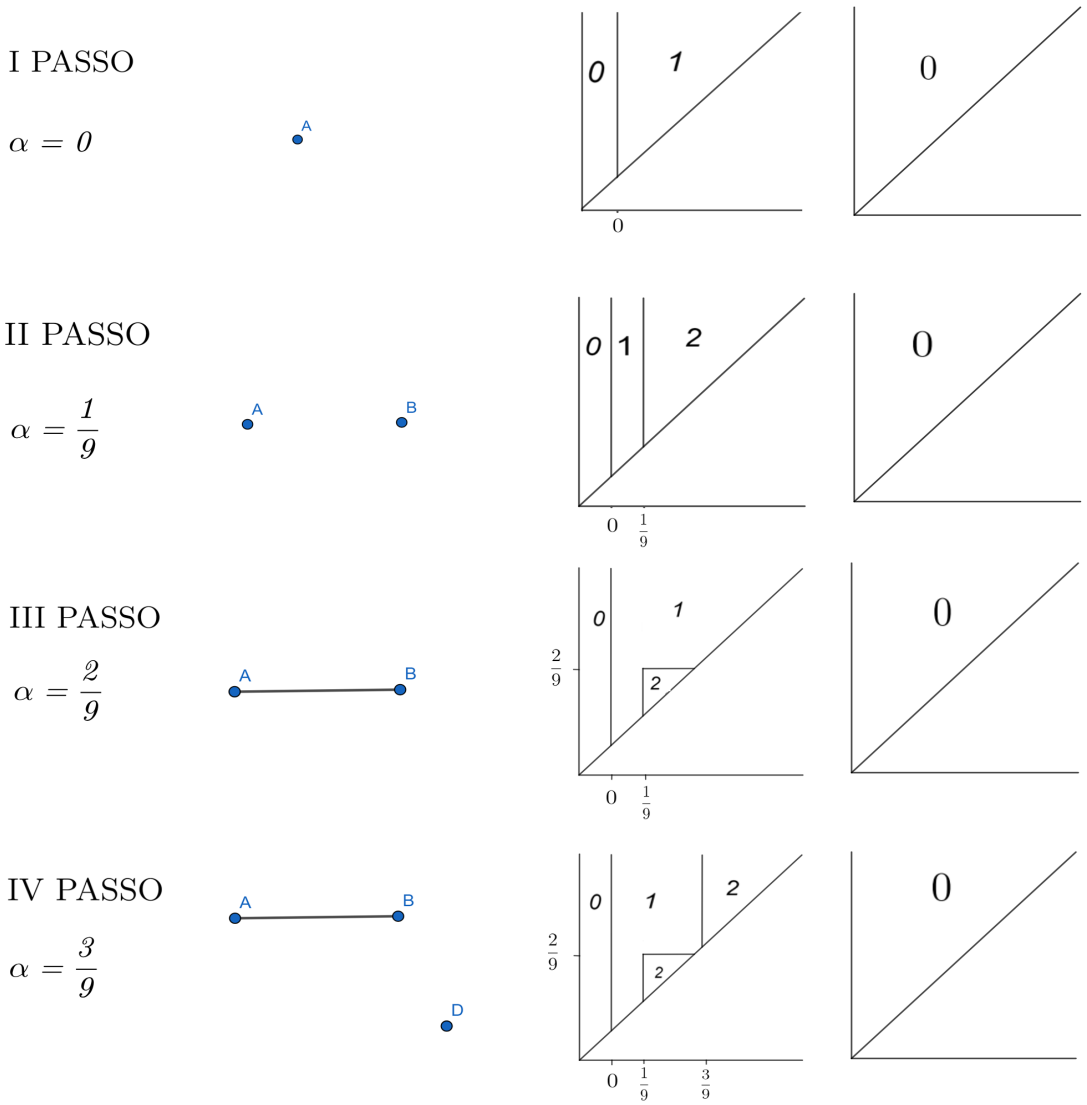
Definizione 2.10 (**Filtrazione**).
Per un dato parametro $\alpha \in [0, 1]$ si definisce la filtrazione $\mathcal{L}$ come una collezione di complessi simpliciali L_α tali che per ogni sequenza ordinata $0 = \alpha_0 < \alpha_1 < \ldots < \alpha_m = 1$ vale $\emptyset = L_{\alpha_0} \subseteq L_{\alpha_1} \subseteq \ldots \subseteq L_{\alpha_m} = L$.

Il minimo valore di α in cui un simplesso diventa un elemento di L_α si può considerare come tempo di nascita del simplesso.
Dato un dissimilarity network D_X^K si costruisce una filtrazione $\mathcal{L}(D_X^K)$ considerando il valore di dissimilarità $r_X^k(x_{0:k})$ come tempo di nascita del simplesso $[x_{0:k}]$, cioè

$$[x_{0:k}] \in L_\alpha \Leftrightarrow r_X^k(x_{0:k}) \leq \alpha \tag{2.7}$$

Proposizione 2.2.1. *$\mathcal{L}(D_X^K)$ costruita secondo la relazione (2.7) è una filtrazione ben definita.*

Possiamo ora considerare la filtrazione come una collezione di insiemi di sottolivello, e l'avanzare del tempo come funzione filtrante. Allora, come abbiamo già visto nella Figura 1.7, possiamo andare ad analizzare la comparsa e la scomparsa delle caratteristiche omologiche nella sequenza di complessi simpliciali che porta alla costruzione del dissimilarity network.
La Figura 2.4 mostra a sinistra la realizzazione della filtrazione del dissimilarity network dell'esempio precedente sulla comunità di ricerca, a destra le corrispondenti 0-PBN e 1-PBN.



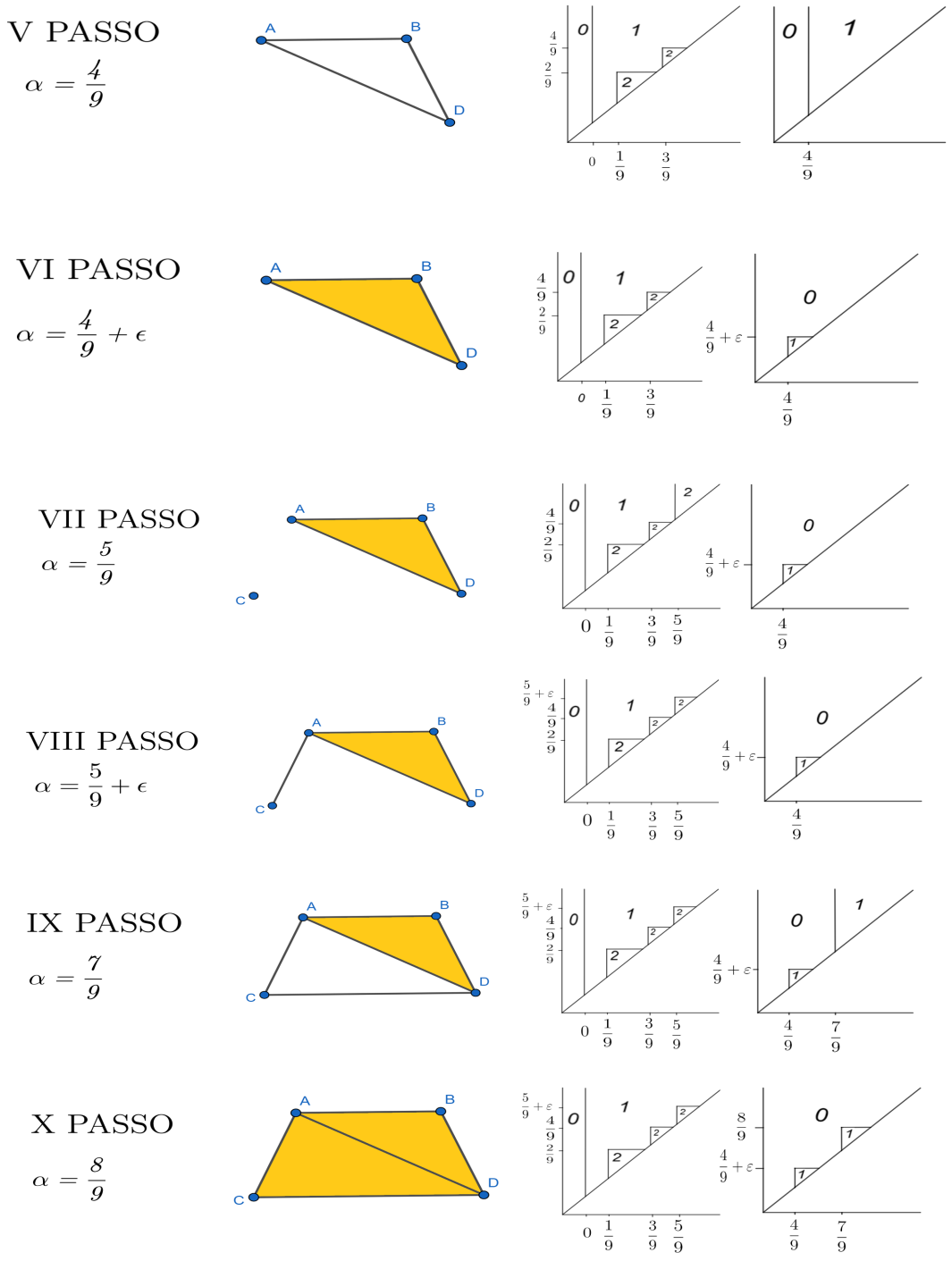
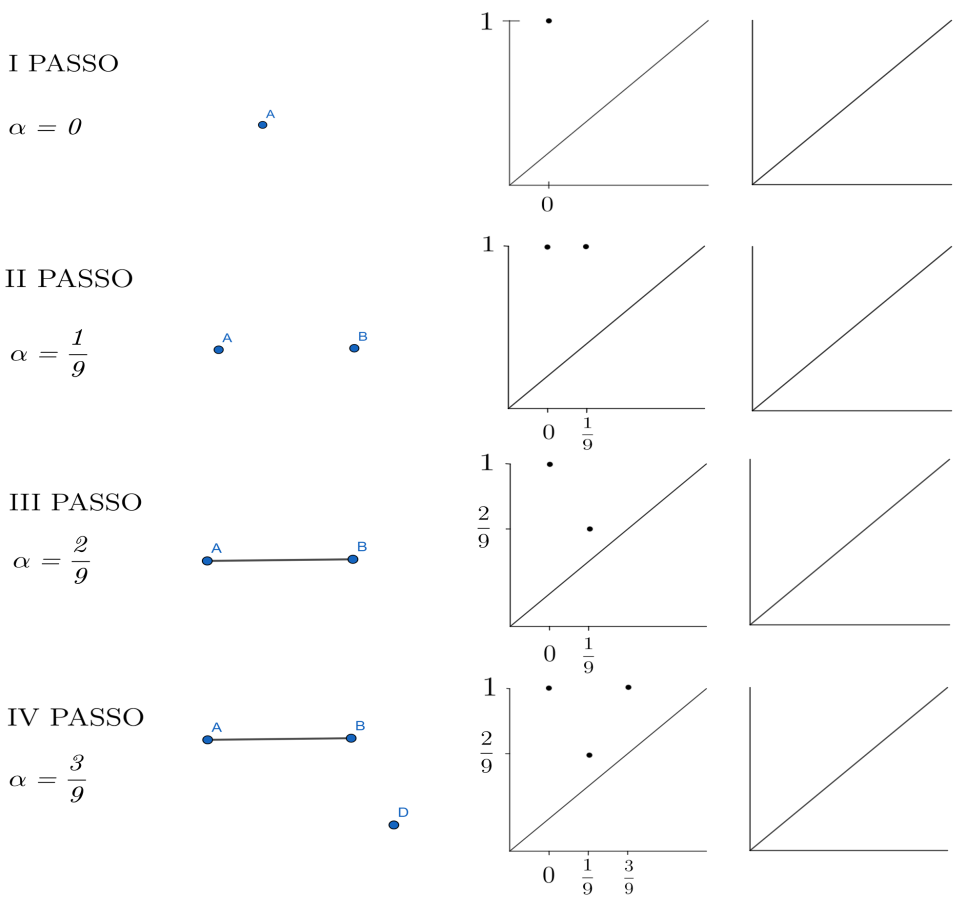


Figura 2.4: Costruzione della filtrazione con le corrispondenti 0-PBN e 1-PBN

Proposizione 2.2.2. *Dato un dissimilarity network D_X^K, qualunque sia il grado k, ogni dissimilarità $r_X^k(x_{0:k})$ tra vettori con elementi unici appare o al tempo di morte delle caratteristiche omologiche $(k-1)$-dimensionali o al tempo di nascita di quelle k-dimensionali.*

Questa proposizione lascia intendere che differenti dissimilarity network generano differenti diagrammi di persistenza. Ragionevolmente possiamo allora usarli come strumenti per la discriminazione delle reti. A tal proposito risulta allora fondamentale il Teorema 1.2.1.
Quando andiamo a comparare due high order networks attraverso i rispettivi diagrammi di persistenza, sapere che la distanza di matching limita inferiormente la pseudodistanza naturale ci fornisce una importante considerazione: una differenza considerevole tra le caratteristiche omologiche persistenti lascia intendere che troveremo una altrettanto considerevole divergenza anche tra i network.
Andiamo ad analizzare questo fatto costruendo i diagrammi di persistenza dell'esempio precedente.



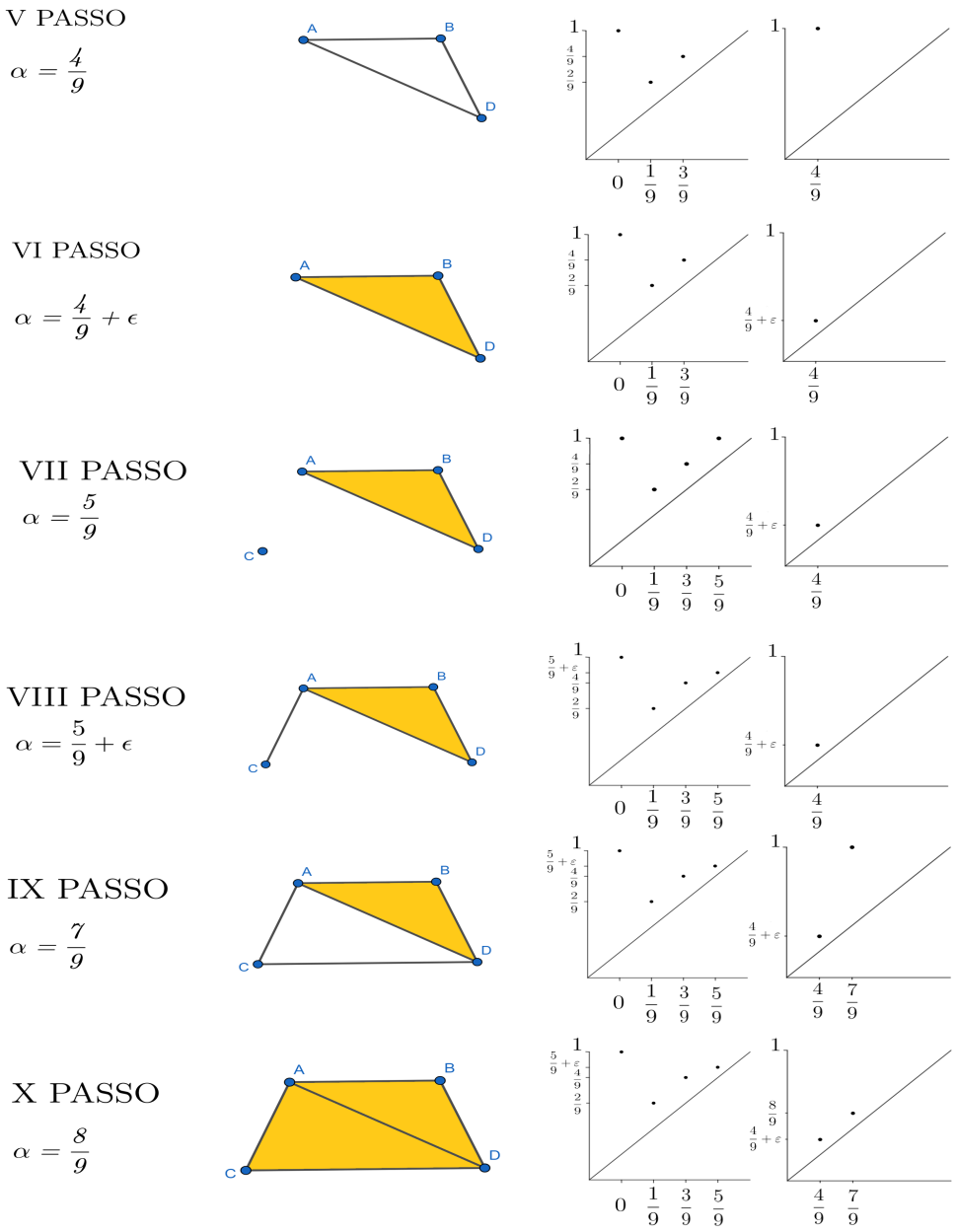


Figura 2.5: Costruzione della filtrazione con i corrispondenti diagrammi di persistenza

Consideriamo prima i diagrammi di persistenza 0-dimensionali. Da quanto si differenziano tra loro possiamo ottenere importanti informazioni su quanto divergono i relativi complessi simpliciali.
Per esempio i diagrammi (V) e (VI) sono uguali quindi, parlando di componenti connesse, i relativi network risultano essere molto simili. Anche il diagramma (VIII) si differenzia di poco da (V): è come se considerassimo il punto (C) come un prolungamento del punto (A), influisce quindi poco sulle caratteristiche omologiche 0-dimensionali. Lo stesso discorso vale per (IX) e (X) in relazione a (V).
Per quanto riguarda i diagrammi di persistenza 1-dimensionali possiamo notare che la distanza di matching maggiore si ottiene se prendiamo in esame (IX) e (V). Simili tra loro ma considerevolmente diversi dagli altri diagrammi, infatti i complessi simpliciali corrispondenti sono gli unici a presentare una caratteristica omologica 1-dimensionale, il buco.
Facciamo ora una considerazione sull'implementazione computazionale di questi passaggi, senza entrare nei particolari delle procedure algoritmiche che determinano filtrazioni e diagrammi di persistenza. Per alleggerire la complessità di calcolo possiamo ridurre il numero di punti di un diagramma di persistenza rimuovendo quelli molto vicini alla diagonale. Questo accorgimento è giustificato perché questi punti rappresentano effimere caratteristiche omologiche, come il buco che nasce in (V) al tempo $4/9$ e muore subito dopo al tempo $4/9 + \varepsilon$.
(Questi punti sono solitamente generati da "rumore" nelle osservazioni.)

2.3 Proximity networks

Nei dissimilarity network le funzioni di correlazione $r_X^k(x_{0:k})$ descrivono quanto sono diversi o lontani tra loro gli elementi del vettore. Alternativamente possiamo andare a considerare delle relazioni che esprimono invece un livello di similarità, questo è il caso dei proximity network. La disuguaglianza definita in (2.6) diventa allora

$$r_X^k(x_{0:k}) \leq r_X^{k-1}(x_{0,k-1}) \tag{2.8}$$

con l'uguaglianza che si ottiene se e solo se il k-esimo elemento di $x_{0:k}$ appare anche in $x_{0,k-1}$.
Commentiamo subito l'esempio in Figura 2.6 che mostra un diverso aspetto del gruppo di coautori: ora le relazioni r di ordine k descrivono il numero di pubblicazioni su cui hanno lavorato insieme gli elementi di un dato vettore. In particolare r_X^0 ci dice quante pubblicazioni ha un autore, tenendo conto sia di quando ha lavorato da solo sia di quando ha collaborato con altre persone. Questo numero è normalizzato secondo il totale degli articoli presi in considerazione nel network, 19 nel nostro caso. Consideriamo l'autore A: ha pubblicato 11 articoli in totale, di cui 4 sono scritti insieme a B, 2 insieme a D e 2 insieme a C. Tra questi 2 sono scritti dal terzetto A,B,D e un articolo da A,B,C. È evidente che $r_X^1(x,y) \leq r_X^0(x)$ e $r_X^1(x,y) \leq r_X^0(y)$, oppure $r_X^2(x,y,z) \leq r_X^1(y,z)$ e via dicendo.

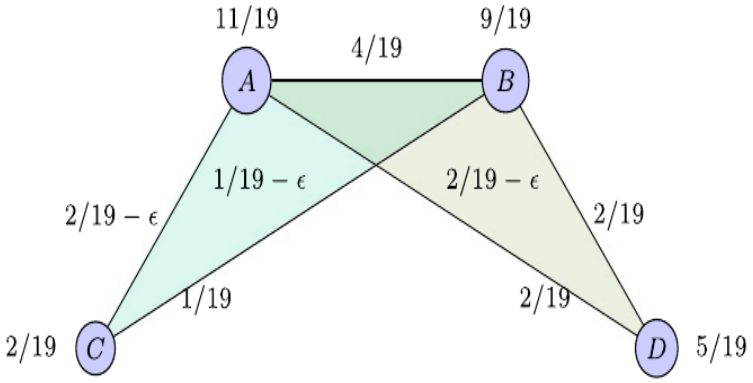


Figura 2.6: Un proximity network.

Per ogni proximity network P_X^K con funzioni di relazione $r_X^k(x_{0:k})$ si può definire il duale dissimilarity network D_X^K con funzioni di relazione $\tilde{r}_X^k(x_{0:k}) := 1 - r_X^k(x_{0:k})$. Segue che tutte le considerazioni fatte nella sezione precedente valgono anche con queste reti. Possiamo cioè costruire a partire dalla (2.8) una filtrazione ben definita, e da lì possiamo analizzare le differenze tra i proximity network andando a calcolare la distanza di matching dai rispettivi diagrammi di persistenza.
Ripercorrendo la ricerca fatta in [7] si procede ora con lo studio di esistenti reti di coautori che rispecchiano il modello di proximity network appena esaminato in cui le funzioni

di relazione indicano il numero di pubblicazioni di un singolo autore, di una coppia o di un terzetto.
Vengono prese in analisi 5 riviste dalla comunità matematica e 6 riviste da quella degli ingegneri. Per ogni rivista vengono presi in considerazione due lustri, dal 2004 al 2008 e dal 2009 al 2013. Per tre particolati riviste delle 6 di ingegneria (TAC, TSP, TWC) vengono costruiti i network di ogni singolo anno dal 2004 al 2013. Intuendo che le reti costruite a partire dalla stessa comunità di ricerca o provenienti dalla stessa rivista possano avere simili modelli di collaborazione, si mostra di seguito che la distanza di matching riesce ad identificare questi schemi rivelandosi un ottimo strumento per distinguere le reti che provengono da comunità scientifiche con interessi diversi.
Prima di procedere introduciamo lo scaling multidimensionale: è una tecnica che produce visualizzazioni in 2 o 3 dimensioni di similarità tra dataset formati da elementi multidimensionali o distanze tra punti. I grafici ottenuti si chiamano embedding. Nel processo di creazione di queste visualizzazioni vi è un'inevitabile perdita di informazione.

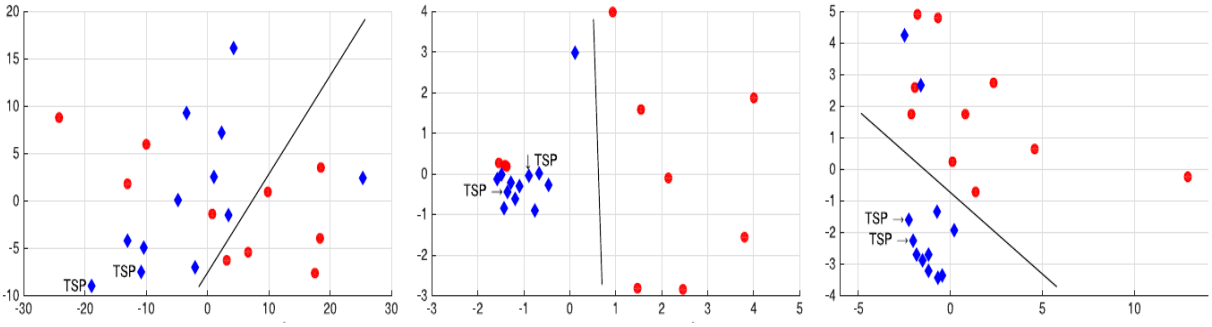


Figura 2.7: Embeddings rispetto b_0, b_1, b_2.

La Figura 2.7 mostra gli embeddings sul piano euclideo calcolati rispetto le distanze di matching 0-dimensionali b_0 a sinistra, b_1 al centro e b_2 a destra. Ogni rivista viene quindi identificata da due punti, uno per ogni lustro, che vengono posizionati sul piano cercando di rispettare al meglio la metrica imposta dalle distanze b_k calcolate tra ogni elemento del dataset (22 riviste). Le 12 reti degli ingegneri (sei per ogni lasso di tempo considerato) sono rappresentate in blu, le 10 reti dei matematici sono invece i punti rossi. Il clustering, fenomeno di raggruppamento tra elementi simili, non è molto chiaro in b_0; in b_1 e b_2 invece i punti blu sono distintamente separati da quelli rossi. In ognuno dei tre grafici è tracciata una retta che funge da separatore tra i due gruppi. Questo confine rende manifesta la presenza degli outliers, valori che possono essere considerati anomali, distanti dalle osservazioni. Bisogna prestare molta attenzione agli outlier perché influiscono sull'analisi del modello e sulla sua capacità predittiva. Spesso questi valori anomali sono l'obiettivo della ricerca, esistono infatti problemi di predictive analytics categoriz-

zati come problemi di anomaly detection.
Tornando ad analizzare la Figura 2.7 notiamo che le reti costruite a partire dalla stessa rivista tendono ad essere abbastanza vicine. Sui grafici è riportato l'esempio della rivista Trans. Signal Processing (TSP): è chiaro che le loro differenze nelle omologie sono considerevolmente basse. Considerando che le reti delle comunità di ingegneri hanno, in generale, caratteristiche omologiche 0 dimensionali che nascono presto e muoiono tardi, mentre le caratteristiche di dimensione maggiore nascono tardi e muoiono conseguentemente ancora più in là nel tempo, si è data la seguente interpretazione: in queste comunità esistono piccoli gruppi che collaborano poco tra di loro, cioè è comune trovare una forte collaborazione tra coppie di autori piuttosto che in collettivi di 3 o più persone. Questo processo di discriminazione basato sulla distanza di matching ha successo anche nell'identificare all'interno della comunità degli ingegneri i gruppi con differenti interessi di ricerca. Nella Figura 2.8 vengono considerate le reti costruite anno per anno con le pubblicazioni delle riviste Trans. Automatic Control (TAC), Trans. Wireless Communication (TWC) e Trans. Signal Processing (TSP).

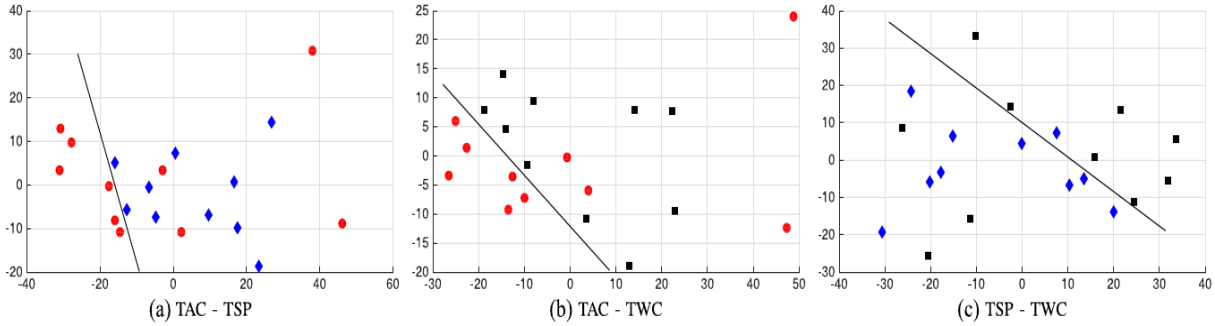


Figura 2.8: Embeddings delle comunità di ingegneri con differenti interessi di ricerca.

Capitolo 3

Reti di collaborazione che evolvono nel tempo

È interessante vedere come in [10] siano giunti alle stesse conclusioni percorrendo una strada apparentemente molto diversa. Nell'esempio precedente sulle pubblicazioni delle tre riviste ingegneristiche, per ogni anno dal 2004 al 2013 e per ognuna di esse, il gruppo di coautori viene considerato come una rete a sé stante; vengono cioè assegnati ad ogni rivista 10 diagrammi di persistenza indipendenti tra loro. Il diverso approccio che andiamo ora ad analizzare avrebbe invece considerato le 10 reti come passaggi all'interno di una filtrazione di un'unica rete di collaborazione che cresce secondo una parametrizzazione temporale.
Formalmente si ha una sequenza di network $\{\mathcal{N}_t,\ t = 0, 1, \ldots, T\}$ in cui ad ogni istante t vengono aggiunti nuovi collaboratori. Questa sequenza viene indicata come filtrazione temporale.
L'articolo si sviluppa in questo modo: prese in considerazione due reti, quella di autori informatici che hanno pubblicato insieme e quella di attori che hanno recitato insieme, viene fatta prima un'analisi topologica tradizionale per cercare proprietà e tendenze, e successivamente vengono analizzate le caratteristiche omologiche persistenti partendo dalla definizione di una distanza diversa da quella che abbiamo utilizzato fino ad ora. Procediamo quindi con lo studio della prima parte.
Per ogni rete vengono considerate finestre temporali di 10 anni che non si sovrappongono: per la rete di coautori informatici abbiamo sei finestre, da D_1 che rappresenta il segmento 1950-59 a D_6 che è quello dal 2000 al 2009; allo stesso modo I_1 rappresenta il decennio 1900-09 per la rete di attori, fino al segmento I_1 che va dal 2000 al 2009.
L'analisi indica che i più recenti segmenti D_4, D_5, D_6 presentano più del 99 per cento di tutti i gruppi di omologia 1-dimensionali, i buchi 2-dimensionali, nella loro componente connessa più estesa (LCC). D_1 ha un solo buco che è nell'LCC, D_2 ne ha 6 di cui 4 nell'LCC e D_3 ne ha 358 di cui 346 si trovano nell'LCC, il 96 per cento.
Per quanto riguarda le reti di attori, a parte I_2 che ha il 97,5 per cento di tutti i buchi

nel proprio LCC, in ogni altra finestra temporale la percentuale è sempre superiore al 99 per cento.
In prima analisi abbiamo quindi che la componente connessa più estesa comprende una più che significante percentuale di gruppi di omologia 1-dimensionali. Questa osservazione empirica richiama i risultati teorici per i grafi di Erdős-Rényi secondo i quali la probabilità delle piccole componenti di non essere dei cicli tende ad 1 portando l'ampiezza del grafo al limite. Ciò fa notare che i grafi di Erdős-Rényi non sono dei buoni modelli per le reti che abbiamo preso in considerazione. Nella Figura 3.1 si vede che il numero di Betti 1-dimensionale cresce rapidamente a partire dagli anni '90 per la rete degli informatici rappresentata in rosso, mentre per la rete degli attori, rappresentata in blu, cresce gradualmente.

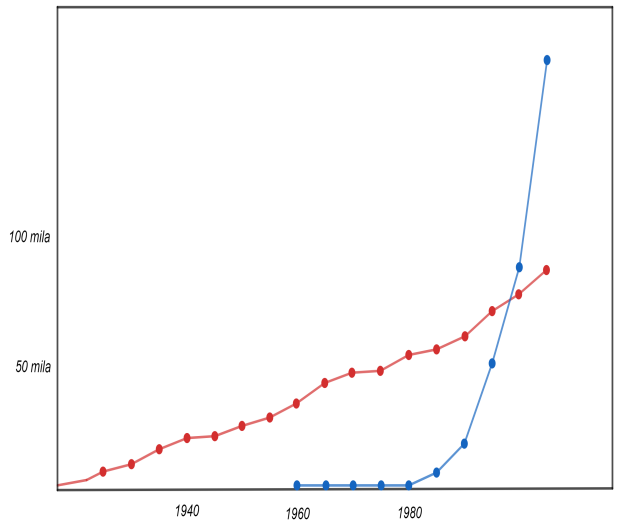


Figura 3.1: Numeri di Betti 1-dimensionali

Equivalentemente la Figura 3.2 mostra che il grado medio decresce dagli anni '40 in poi per la rete di attori mentre cresce costantemente per la rete di autori.

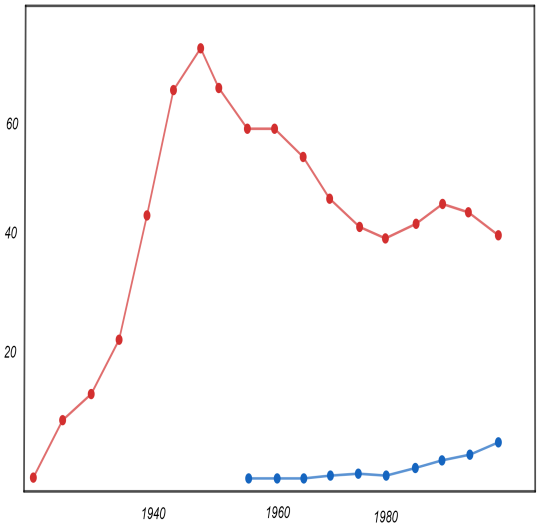


Figura 3.2: Grado medio

Nella Figura 3.3 è rappresentato il numero di cicli per nodo in LCC: cresce costantemente per gli informatici e rimane sostanzialmente invariato per gli attori dagli anni '30 in poi. La figura 3.4 allo stesso modo mostra l'andamento del numero di cicli per bordo. Osserviamo quindi che le proprietà omologiche di entrambe le reti crescono col tempo ma cambiano a ritmi diversi durante i periodi presi in considerazione.

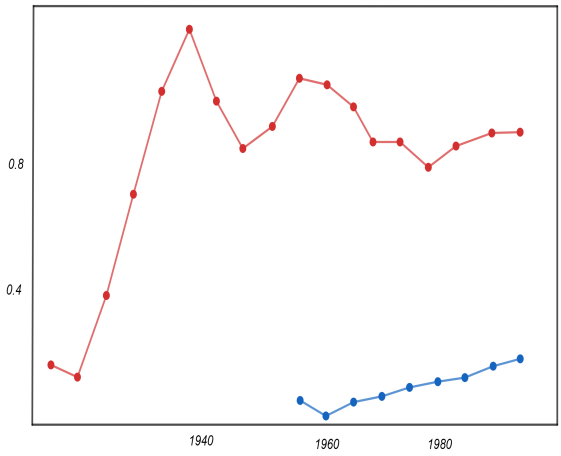


Figura 3.3: Cicli per nodo nell'LCC

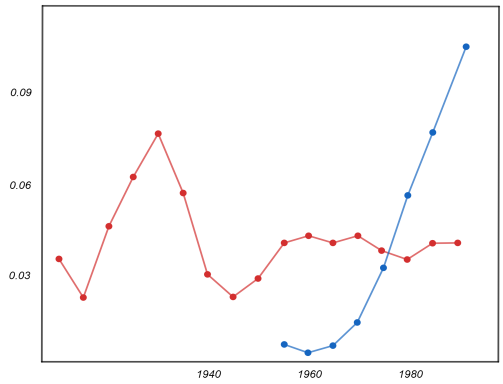


Figura 3.4: Cicli per bordo

Passiamo ora ad una analisi persistente di queste reti.
Al posto della distanza di matching viene introdotta una nuova misura tra reti che crescono nel tempo:
per una sequenza di reti $\{\mathcal{N}_t\ ,\ t = 0, 1, \ldots, T\}$ si definisce una funzione
$g : \{0, 1, \ldots, T\} \to \mathbb{Z}^+$ tale che $g(i)$ rappresenta il numero di nuovi buchi formati al tempo i.
Da questa funzione g si ottiene una funzione di distribuzione cumulativa

$$F(x) = \frac{\sum_{t=0}^{x} g(t)}{\sum_{s=0}^{T} g(s)} \qquad x = 0, \ldots, T$$

e la corrispondente funzione di probabilità f.
Viene poi usata la divergenza di Jensen-Shannon [8] per comparare il ritmo di nascita dei cicli nelle due reti che evolvono nel tempo. Questa misura tra i diversi segmenti di network è riportata graficamente nella Figura 3.5 attraverso lo scaling multidimensionale.
Da questo grafico si nota che i segmenti temporali della rete di attori sono strettamente raggruppati tra loro, distanti da tutti quelli della rete di autori informatici. In questi ultimi invece il clustering è debole: basti guardare quanto sono distanti D_1 e D_2 da tutti gli altri.
Da questi dati si può concludere che mentre la rete degli attori è rimasta topologicamente invariata durante il periodo preso in considerazione, quella degli autori cambia notevolmente dagli anni '70 in poi. Questi valori possono derivare dal fatto che l'industria cinematografica, molto più vecchia rispetto la rete di collaborazioni informatiche, era già matura nel 1950 e le sue proprietà non sono cambiate col passare del tempo. Invece la natura delle collaborazioni si è sviluppata dal 1970 di pari passo con l'innovativo campo scientifico di cui tratta.

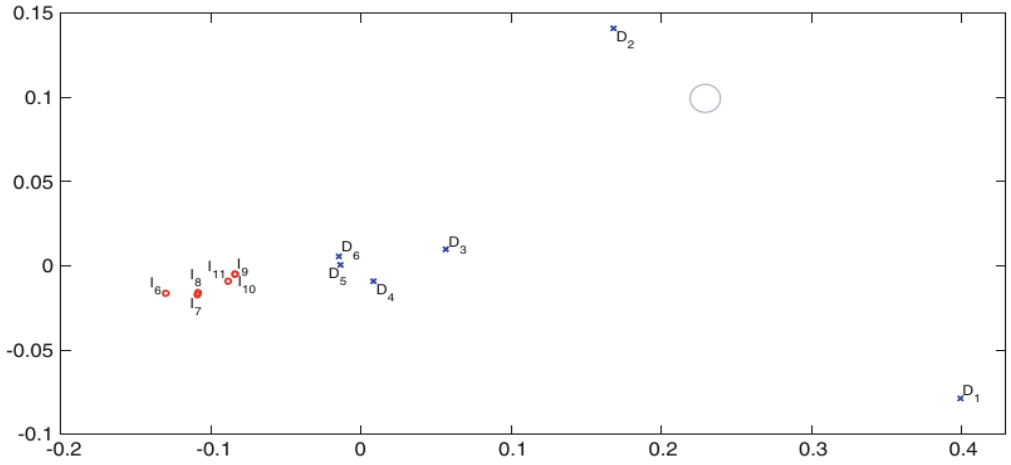


Figura 3.5: Embedding delle distanze di Jensen-Shannon

Capitolo 4

Assortatività del secondo ordine

In quest'ultimo capitolo mettiamo un attimo da parte l'omologia persistente per parlare di una proprietà intrinseca delle reti sociali che si esprime in funzione del numero di connessioni dei nodi, cioè il grado. Stiamo parlando dell'assortatività del secondo ordine.
Quella del primo ordine è la proprietà secondo la quale se due nodi sono connessi allora tendono ad avere lo stesso grado; ciò rispecchia la tendenza delle persone a legare con chi ha un simile livello di importanza sociale. In sociologia questa proprietà è chiamata omofilia.
Se due nodi sono connessi si può andare a misurare il grado di correlazione tra i loro contatti più significativi piuttosto che tra i nodi stessi; questa è l'assortatività del secondo ordine. In [13] vengono studiati diverse reti, sociali e non, per trovare spiegazioni e implicazioni all'esistenza di questa proprietà. Portiamo qui l'esempio di una comunità di attori (in cui due soggetti sono collegati se hanno recitato in uno stesso film), di una comunità di fisici ricercatori (i collegamenti sono dati dalla cooperazione nella pubblicazione di un articolo), di una comunità di musicisti (sono in relazione se hanno suonato con una stessa band) e di diverse reti non sociali riguardanti la biologia e la tecnologia.

Rete	Nodi	Link	r	$\mathcal{R}_{max}$	$\mathcal{R}_{avg}$
Attori	82 mila	3 milioni	0.206	0.813	0.836
Fisici	12mila	40 mila	0.161	0.647	0.68
Musicisti	200	2700	0.02	0.307	0.543
Internet	11 mila	23 mila	-0.195	0.036	-0.097
Proteine	4600	14800	-0.137	0.033	-0.046
Rete casuale	10 mila	30 mila	0	0	0

Nella tabella r è il coefficiente assortativo di primo grado proposto in [9]: si ha $-1 \leq r \leq 1$ dove $r = 1$ indica un perfetto livello assortativo, cioè ogni link connette nodi con lo stesso grado, $r = -1$ al contrario indica un elevato livello disassortativo e $r = 0$ ci dice che non vi è alcuna correlazione tra i gradi.
$\mathcal{R}_{max}$ indica il coefficiente assortativo del secondo ordine calcolato considerando il grado massimo dei nodi mentre $\mathcal{R}_{avg}$ indica il coefficiente calcolato secondo la media dei gradi. Si può notare che le reti sociali mostrano un valore positivo del coefficiente r e un valore considerevolmente più alto dei coefficienti del secondo ordine. Ciò indica che in questi modelli le persone giudicano lo status di altri individui non solo sull'importanza dell'individuo stesso ma, in modo cruciale, sul rilievo sociale dei suoi collaboratori. La differenza maggiore tra di due livelli di assortatività si osserva nella comunità dei musicisti: anche se i soggetti presi in considerazione dimostrano una forte parità sociale questa non può essere rilevata misurando il valore dei musicisti stessi ma attraverso il valore dei colleghi con cui loro hanno performato.
Come era prevedibile, nelle reti in cui viene a mancare la componente umana i coefficienti assortativi del secondo ordine sono sia nulli sia negativi. In [13] si sono chiesti se i forti valori di $\mathcal{R}_{max}$ e $\mathcal{R}_{avg}$ trovati nelle reti sociali possano essere semplicemente derivati dall'incremento dei contatti, visto che un nodo ha sempre più collegamenti del secondo ordine rispetto quelli del primo. Si è calcolata allora l'assortatività di ordini superiori e già i valori dei coefficienti di terzo grado risultano essere decisamente più piccoli.
Spesso la struttura della rete è dominata da nodi centrali estremamente ben connessi. Rimuovendo questo nucleo e ricalcolando i coefficienti si nota che per alcune reti l'assortatività del secondo ordine diventa persino più forte. Quindi questa proprietà non può essere collegata alla presenza di nodi con un grado molto elevato. Non si è trovata una correlazione neanche prendendo in esame il coefficiente di clustering ed altri strumenti, tra i quali viene a mancare la topologia persistente.
Potrebbe essere allora interessante lavorare su questo fronte: per esempio prendiamo il caso in cui il contatto più rilevante per entrambi i nodi di un link sia la stessa persona, si verrebbe a formare cioè un triangolo. L'idea allora è cercare nelle reti sociali un collegamento tra rilevanti caratteristiche omologiche 1-dimensionali (o maggiori) con un alto coefficiente assortativo del secondo ordine.
Nell'esempio del Capitolo 2 la comunità di ingegneri, a differenza della comunità di matematici, ha caratteristiche omologiche 1-dimensionali che nascono tardi e hanno vita breve. Se penso alla comunità di fisici di questo capitolo, che presenta un alto valore assortativo, mi viene da pensare che si potrebbe trovare un'assortatività del secondo ordine del gruppo degli ingegneri più bassa rispetto quella del gruppo di matematici.

Capitolo 5

Conclusioni

Nel corso dei capitoli precedenti si è quindi osservato quanto gli strumenti dell'omologia persistente risultino efficaci nei processi di classificazione e differenziazione delle reti complesse.
All'inizio vengono rappresentate molto intuitivamente come ipergrafi pesati e da qui il passaggio ai complessi simpliciali è semplice. I valori di ogni simplesso ben definiscono una filtrazione nei dissimilarity e proximity network, e in questa siamo andati a vedere quando nascono, muoiono ma soprattutto persistono le caratteristiche omologiche della struttura. Siccome la differenza tra due network è approssimabile attraverso la distanza di matching siamo andati a ricavare quest'ultima dai diagrammi di persistenza per giungere infine, per mezzo dello scaling multidimensionale, ad una rappresentazione bidimensionale dell'insieme di reti che rispetti le misure trovate, che renda quindi manifeste le difformità, o le conformità, tra i network.
Nel terzo capitolo si è visto parallelamente come anche la topologia non persistente possa dare contributi nel processo di discriminazione delle reti complesse. Nell'ultimo capitolo invece si è cercato un collegamento tra l'omologia e l'assortatività del secondo ordine nelle reti sociali.

Capitolo 6

Appendice

6.1 Embeddings

Nell'elaborato si è introdotto il concetto di multidimentional scaling (MDS), cioè una tecnica che, partendo dalla conoscenza delle distanze tra tutte le coppie di punti di un qualsiasi insieme, sceglie uno spazio metrico delle adatte dimensioni e vi colloca i punti in modo da rispettare nel miglior modo possibile le distanze conosciute. Per esempio nel capitolo 2 si sono visualizzate le reti di ingegneri e matematici su un piano euclideo rispettando al meglio le distanze di matching tra ogni coppia di riviste. Le distanze date come input si dicono dissimilarità e l'insieme di coordinate che l'algoritmo produce in un prestabilito spazio metrico si dice configurazione degli output. Per ottenere la migliore configurazione si cerca il minimo globale di una data funzione di ottimizzazione, cioè l'errore di embedding o misura dello stress. Tradizionalmente MDS usa il piano euclideo come spazio per le visualizzazioni. È interessante vedere come in [2] si analizza invece la convenienza nell'uso del piano iperbolico, spazio che ottiene le configurazioni che codificano gli input iniziali con la minor perdita di informazioni.
Come modello per il piano iperbolico si è usato il disco di Poincaré (PD). Parleremo quindi in seguito di PD-MDS per indicare l'algoritmo. Come misura dello stress si è usata la funzione Sammon proposta in [11]: rappresenta la sommatoria normalizzata del quadrato della differenza tra le dissimilarità originali δ_{jk} e la distanza d_{jk} ottenuta dalla configurazione degli output:

$$E = c \sum_{j=1}^{n} \sum_{k=j+1}^{n} c_{jk}(d_{jk} - a\delta_{jk})^2$$

. È una formula generale da cui si possono ricavare diverse funzioni per l'errore di embedding sostituendo alle costanti c, c_{jk} e a degli appropriati valori.

Trattando del piano iperbolico la distanza d_{jk} sarà data da:

$$d_{\mathbb{D}}(z_j, z_k) = 2atanh\frac{|z_j - z_k|}{|1 - z_j\overline{z_k}|}$$

dove $\mathbb{D} = \{z \in \mathbb{C} || z| < 1\}$ è il disco di Poincaré.
Con questa distanza iperbolica la funzione E manca di invarianza di scala, di cui a è il fattore. L'assenza di questa proprietà conferisce un grado di libertà in più nell'ottimizzazione dell'errore di embedding, cioè proprio il fattore a.
Nell'articolo si studia quindi la dipendenza dell'errore da questo fattore. L'esperimento riportato si basa su grafi non reali generati in maniera casuale i cui dati di input vivono su superfici con curvatura positiva, zero o negativa, cioè rispettivamente la sfera, il piano euclideo e il piano iperbolico (riporterò solo questi ultimi due esempi). Le corrispondenti distanze tra ogni coppia di nodi sono come abbiamo visto le dissimilarità. Si sono anche considerati input perturbati ottenuti rimpiazzando ogni δ_{jk} con un valore scelto uniformemente nell'intervallo $[(1 - e_m)\delta_{jk}, (1 + e_m)\delta_{jk}]$ per un dato livello di rumore $e_m < 1$.
I risultati vengono riportati nelle Figure 6.1 e 6.2. I grafici (c) ed (e) illustrano la variazione dell'errore di embedding per i dati di input non perturbati, con il numero di nodi come parametro, 20 e 60 punti. I grafici (d) e (f) illustrano invece la variazione dell'errore dei dati perturbati e sono parametrizzati secondo l'ammontare dell'errore, $e_m = 0, 10, 20, 30\%$.

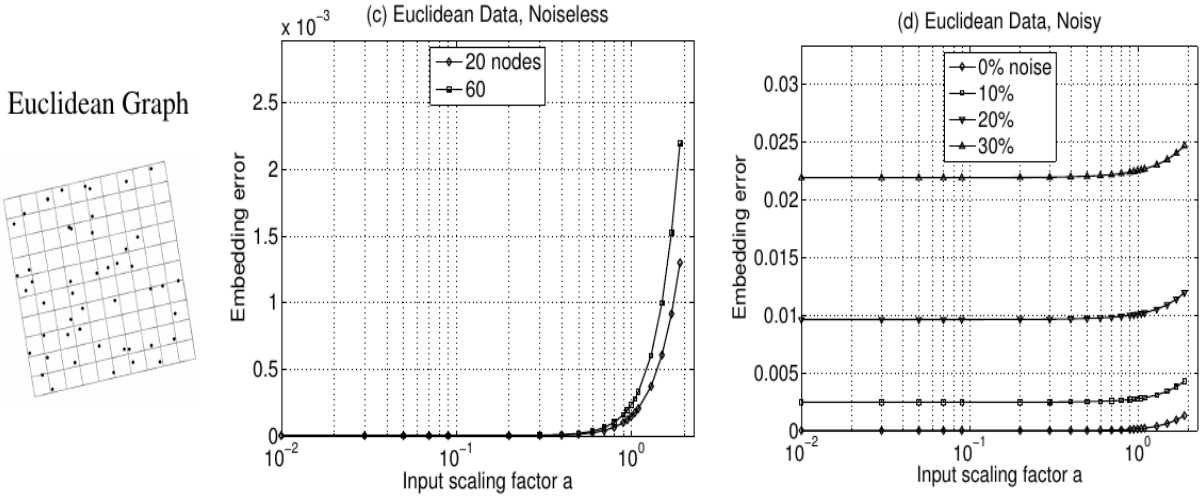


Figura 6.1: Come varia l'errore di embedding rispetto il fattore di scala a quando si immerge un grafo che vive nel piano euclideo sul disco di Poincaré usando PD-MDS.

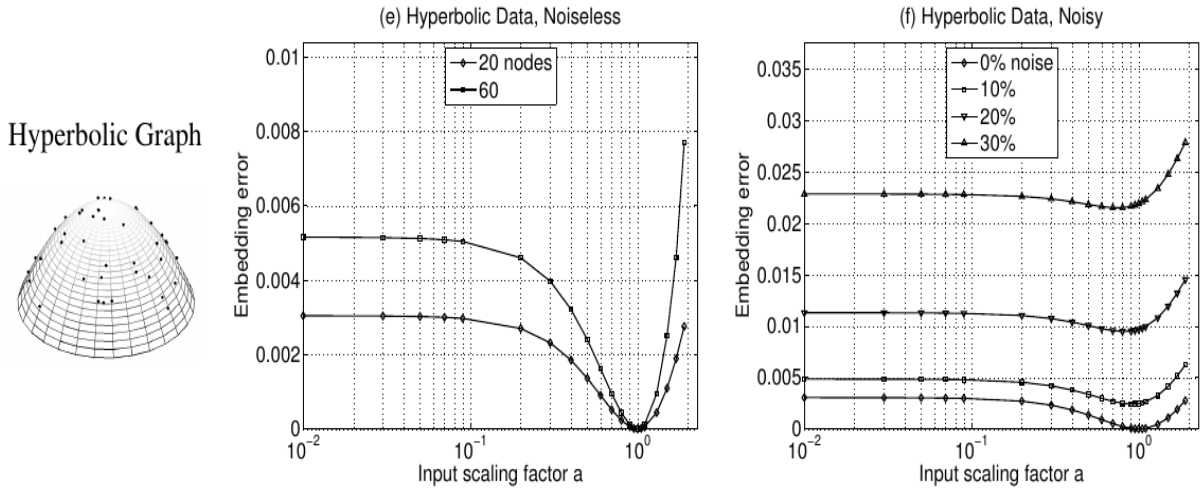


Figura 6.2: Come varia l'errore di embedding rispetto il fattore di scala a quando si immerge un grafo che vive nel piano iperbolico sul disco di Poincaré usando PD-MDS.

Bibliografia

[1] Gunnar Carlsson. «Topology and data». In: *Bulletin of the American Mathematical Society* 46.2 (2009), pp. 255–308.

[2] Andrej Cvetkovski e Mark Crovella. «Low-stress data embedding in the hyperbolic plane using multidimensional scaling». In: *Appl. Math* 11.1 (2017), pp. 5–12.

[3] Massimo Ferri. «Persistent topology for natural data analysis– A survey». In: *Towards Integrative Machine Learning and Knowledge Extraction*. Springer, 2017, pp. 117–133.

[4] Massimo Ferri. «Why topology for machine learning and knowledge extraction?» In: *Machine Learning and Knowledge Extraction* 1.1 (2018), pp. 115–120.

[5] Patrizio Frosini. «Measuring shapes by size functions». In: *Intelligent Robots and Computer Vision X: Algorithms and Techniques*. Vol. 1607. International Society for Optics e Photonics. 1992, pp. 122–134.

[6] Weiyu Huang e Alejandro Ribeiro. «Metrics in the space of high order networks». In: *IEEE Transactions on Signal Processing* 64.3 (2016), pp. 615–629.

[7] Weiyu Huang e Alejandro Ribeiro. «Persistent homology lower bounds on high-order network distances». In: *IEEE Transactions on Signal Processing* 65.2 (2017), pp. 319–334.

[8] J Lin e SKM Wong. «A new directed divergence measure and its characterization». In: *International Journal Of General System* 17.1 (1990), pp. 73–81.

[9] Mark EJ Newman. «Assortative mixing in networks». In: *Physical review letters* 89.20 (2002), p. 208701.

[10] Siddharth Pal et al. «Comparative topological signatures of growing collaboration networks». In: *International Workshop on Complex Networks*. Springer. 2017, pp. 201–209.

[11] John W Sammon. «A nonlinear mapping for data structure analysis». In: *IEEE Transactions on computers* 100.5 (1969), pp. 401–409.

[12] Alessandro Verri et al. «On the use of size functions for shape analysis». In: *Biological cybernetics* 70.2 (1993), pp. 99–107.

[13] Shi Zhou, Ingemar J Cox e Lars K Hansen. «Second-order assortative mixing in social networks». In: *International Workshop on Complex Networks*. Springer. 2017, pp. 3–15.