19/06/2015

Metodi matematici per il calcolo di grandezze caratteristiche nell'analisi di reti complesse

Carlo Sau

Sommario

1. Introduzione

Al giorno d'oggi diversi campi applicativi richiedono l'analisi di reti molto grandi. Le reti sono infatti largamente utilizzate per l'implementazione di sistemi complessi o per la loro efficiente modellazione. Esempi di implementazione di reti vaste possono essere il sistema di produzione e distribuzione dell'energia elettrica a livello nazionale oppure quello relativo alla telefonia mobile e fissa. Casi in cui invece le reti sono impiegate come potenti modelli per l'analisi di sistemi complessi possono essere quello della rappresentazione delle interazioni tra proteine in corrispondenza di eventi biochimici e/o forze elettrostatiche o ancora quello delle relazioni di conoscenza tra utenti nei social network. Tutti gli esempi citati si riferiscono a contesti molto estesi che, in termini di rete, si traducono in un elevato numero di nodi e connessioni. Per tale motivo, anche delle operazioni molto semplici su tali reti possono essere critiche dal punto di vista computazionale, in particolare per latenza e memoria richiesta.

Grandezze tipicamente di interesse nell'analisi di una rete sono l'importanza di un nodo, strettamente connessa al numero di altri nodi raggiungibili dallo stesso, e la facilità di comunicazione tra due nodi, legata invece al numero di connessioni frapposte tra tali nodi. Il calcolo di queste e altre grandezze di interesse si riduce alla risoluzione di espressioni del tipo $\boldsymbol{u}^T f(A)\boldsymbol{w}$, dove A è la matrice di adiacenza che rappresenta la rete in questione, f è una funzione non lineare (ad esempio la funzione esponenziale) e $\boldsymbol{u}$ e $\boldsymbol{w}$ sono dei vettori (tipicamente versori). Nella maggior parte delle applicazioni reali non è necessario calcolare le grandezze in questione per tutti i nodi della rete, ma il problema si focalizza solo sui nodi che presentano i valori più elevati. Per tale motivo, in presenza di reti vaste, e dunque di A molto grandi, il calcolo completo della grandezza su tutta la rete risulta poco efficiente.

Per matrici A simmetriche, è possibile utilizzare metodi alternativi quali la fattorizzazione spettrale parziale e le regole di quadratura per focalizzare il problema solamente sulla sotto porzione di rete di interesse. Per matrici non simmetriche sono comunque applicabili metodi analoghi con alcuni accorgimenti. Entrambi i metodi forniscono dei limiti inferiore e superiore per il valore di $\boldsymbol{u}^T f(A)\boldsymbol{w}$. Tuttavia il primo metodo, la fattorizzazione spettrale parziale, riesce a individuare velocemente i valori più elevati ma ne fornisce dei limiti poco precisi mentre il secondo, le regole di quadratura, fornisce dei limiti più precisi del primo ma ha un carico computazionale maggiore. Una combinazione di tali metodi, attuata in modo tale da sfruttare i punti di forza di entrambi, può portare a una configurazione ottimale per la risoluzione del problema. In particolare la fattorizzazione spettrale può essere utilizzata per individuare velocemente i valori di interesse, mentre le regole di quadratura possono essere adottate in un secondo momento sui valori individuati precedentemente per affinarne la stima.

2. Generalità e principali metriche delle reti

Una rete è un insieme di entità, dette nodi, esattamente identiche (rete omogenea) oppure con diversa funzionalità e/o struttura (rete eterogenea). I nodi sono connessi tra loro attraverso degli archi che possono essere orientati o meno e possono essere pesati. L'orientamento indica il verso di percorrenza di un arco mentre il peso ne indica la facilità di percorrenza. Si definisce cammino (*walk*) all'interno della rete una sequenza ordinata di nodi in cui nodi adiacenti sono connessi da un arco nella rete. Una rete viene rappresentata matematicamente attraverso quella che viene chiamata matrice di adiacenza A. Tale matrice ha n righe e n colonne, con n numero di nodi della rete. Il valore degli elementi di A_{ij} è definito come:

$$A_{ij} = \begin{cases} 1, & se\ c'è\ un\ arco\ tra\ i\ nodi\ i\ e\ j \\ 0, & altrimenti \end{cases} \qquad 2.1$$

Dalla matrice di adiacenza così definita, dato $m \geq 1$, è possibile ricavare i cammini di lunghezza m che iniziano dal nodo i e terminano a quello j attraverso l'elemento $[A^m]_{ij}$.

Si consideri la funzione matriciale f:

$$f(A) = \sum_{c=0}^{\infty} c_m A^m, \qquad 2.2$$

in cui i coefficienti c_m sono non negativi (la sommatoria converge) e $c_0 A^0$ è aggiunto per completezza ma non ha un particolare significato. Allora il termine $[f(A)]_{ij}$, con $i \neq j$, fornisce una misura sulla facilità di comunicazione tra i nodi i e j, mentre $[f(A)]_{ii}$ può quantificare l'importanza del nodo i. Funzioni derivate da $f(A)$ sono adottate nella pratica per analizzare reti complesse. Tipicamente, si opta per coefficienti c_m non crescenti all'aumentare di m, in quanto il più delle volte i cammini più corti sono da prediligere rispetto a quelli più lunghi. Una scelta comune è avere dei coefficienti $c_m = \frac{1}{m!}$ o, in altri termini, f è la funzione esponenziale matriciale:

$$f(A) = \exp(A). \qquad 2.3$$

Dalla matrice A così definita è dunque possibile, data una funzione f, caratterizzare la rete a cui essa è riferita. In particolare le misure di maggior interesse sono:

- il *degree* (o grado) del nodo *i*, ottenuto come $[Ac]_i$ con $c = [1, 1, \dots, 1] \in R^n$, indice dell'importanza del nodo;
- la *f-subgraph centrality* del nodo i, ottenuta attraverso $[f(A)]_{ii}$, anch'essa quantificazione dell'importanza del nodo, più sofisticata del *degree*;
- la *f-communicability* tra i nodi i e j, data da $[f(A)]_{ij}$, misura della facilità di comunicazione tra i due nodi.

Entrambe le misure di *f-subgraph centrality* e di *f-communicability* si riducono al calcolo di un'espressione del tipo:

$$\boldsymbol{u}^T f(A) \boldsymbol{w}, \qquad 2.4$$

con $\boldsymbol{u}$ e $\boldsymbol{w}$ versori $e_j = [0, 0, \dots, 0, 1, 0, \dots, 0] \in R^n$ con indice j uguale (*f-subgraph centrality*) o differente (*f-communicability*). Anche altre misure caratteristiche dell'analisi delle reti, quali ad esempio la *f-starting convenience* e la *f-ending convenience* sono riconducibili al calcolo di equazioni della stessa forma della 2.4. Per semplicità nel seguito della trattazione la rete considerata sarà caratterizzata da archi non multipli, non orientati e non pesati, senza loop e con nodi omogenei. Tali caratteristiche determinano una matrice A simmetrica. È possibile rilassare tali vincoli e, per la matrice A risultante, individuare problemi e soluzioni analoghe a quelle esposte nel proseguo della trattazione.

Quando si ha a che fare con reti complesse, cioè con un gran numero di nodi, il problema maggiore nel calcolo delle grandezze caratteristiche, e dunque dell'equazione 2.4, è il calcolo diretto di $f(A)$. Questo può

facilmente avere una durata troppo elevata, in relazione all'applicazione di interesse, o richiedere risorse di memorizzazione eccessive per il contesto in cui viene eseguito.

Una possibile soluzione al problema è l'impiego di regole di quadratura di Gauss per approssimare la soluzione esatta. Tale tecnica, che verrà esposta nella sezione 4, permette di stimare un insieme di valori soluzione dell'equazione 2.4 attraverso i relativi limiti inferiore e superiore, forniti dall'applicazione di due diverse regole di quadratura di Gauss. Tuttavia i metodi basati sulle regole di quadratura hanno lo svantaggio di avere una complessità computazionale dipendente dal numero di soluzioni dell'equazione 2.4. Per tale motivo risultano essere vantaggiosi soltanto laddove siano richieste poche soluzioni. Quando si ha bisogno di un numero elevato di valori di $f(A)$, ad esempio nell'individuazione dei nodi con *f-subgraph centrality* più alta, i metodi basati su regole di quadratura possono essere eccessivamente pesanti dal punto di vista computazionale.

Un altro metodo alternativo al calcolo di $f(A)$ è rappresentato dalla fattorizzazione spettrale parziale. Tale metodo, che verrà esposto nella sezione 3, è anch'esso capace di fornire un'approssimazione delle soluzioni dell'equazione 2.4 sempre attraverso una coppia di limiti inferiore e superiore delle stesse. In particolare il metodo si serve di un sottoinsieme delle coppie di autovalore-autovettore dominanti della matrice A. Tuttavia esso non è in grado di fornire delle approssimazioni accurate quanto il metodo basato sulle regole di quadratura di Gauss precedentemente introdotto, né riesce a determinare l'ordine di importanza dei nodi dall'approssimazione delle misure caratteristiche della rete a essi relative.

Entrambi i metodi citati risolvono il problema del calcolo di un sottoinsieme di soluzioni dell'equazione 2.4. Dalle peculiarità dei metodi stessi è possibile intravedere una complementarità che, se sfruttata, può portare ad una soluzione ottimale che raggiunga il miglior compromesso tra precisione del calcolo e carico computazionale. Tale complementarità è stata sfruttata da Fenu et. al [1] attraverso un nuovo metodo ibrido, esposto nella sezione 5, che sfrutta i punti di forza dei due metodi: la fattorizzazione spettrale parziale viene utilizzata per ridurre il campo di ricerca ai nodi con i valori soluzione dell'equazione 2.4 più elevati ed individuarne un'approssimazione grossolana. Dopodiché la stima viene affinata attraverso il metodo basato sulle regole di quadratura, applicato solo sulla porzione di nodi di interesse, individuata al passo precedente.

3. Fattorizzazione Spettrale Parziale

La fattorizzazione spettrale è un tipo di decomposizione matriciale valida quando la matrice $A \in R^{n\times n}$ presa in considerazione è simmetrica. In tal caso essa si può rappresentare come:

$$A = V\Lambda V^T, \qquad 3.1$$

dove $\Lambda = diag(\lambda_1, \lambda_2, \dots, \lambda_n)$ è la matrice $n \times n$ diagonale contenente gli autovalori $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ di A, mentre $\mathrm{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_n]$ è la matrice $n \times n$ contenente gli autovettori di A. Data una funzione matriciale f non negativa e non decrescente in A, si ha:

$$f(A) = f(V\Lambda V^T) = Vf(\Lambda)V^T = \sum_{k=1}^{n} f(\lambda_k)\boldsymbol{v}_k\boldsymbol{v}_k^T. \qquad 3.2$$

Riprendendo in mano il problema dell'equazione 2.4 si ottiene:

$$\boldsymbol{u}^T f(A)\boldsymbol{w} = \boldsymbol{u}^T f(V\Lambda V^T)\boldsymbol{w} = \sum_{k=1}^{n} f(\lambda_k)\tilde{u}_k\tilde{w}_k, \qquad 3.3$$

con $\tilde{u}_k = \boldsymbol{u}^T\boldsymbol{v}_k$ e $\tilde{w}_k = \boldsymbol{v}_k^T\boldsymbol{w}$.

Se le prime $N < n$ coppie di autovalori ed autovettori sono note, la 3.3 si può approssimare come:

$$\boldsymbol{u}^T f(A)\boldsymbol{w} = \sum_{k=1}^{n} f(\lambda_k\tilde{u}_k\tilde{w}_k) \approx \sum_{k=1}^{N} f(\lambda_k)\tilde{u}_k\tilde{w}_k = F_{\boldsymbol{u},\boldsymbol{v}}^{(N)}. \qquad 3.4$$

Calcolando ora l'errore di tale approssimazione si ha:

$$\left|\sum_{k=1}^{n} f(\lambda_k)\tilde{u}_k\tilde{w}_k - F_{\boldsymbol{u},\boldsymbol{v}}^{(N)}\right| = \left|\sum_{k=N+1}^{n} f(\lambda_k)\tilde{u}_k\tilde{w}_k\right|, \qquad 3.5$$

e applicando la diseguaglianza di Cauchy si ottiene:

$$\left|\sum_{k=N+1}^{n} f(\lambda_k)\tilde{u}_k\tilde{w}_k\right| \le f(\lambda_N)\sum_{k=N+1}^{n}|\tilde{u}_k||\tilde{w}_k| \le f(\lambda_N)\sqrt{\sum_{k=N+1}^{n}\tilde{u}_k{}^2}\sqrt{\sum_{k=N+1}^{n}\tilde{w}_k{}^2}$$

$$= f(\lambda_N)\left(1-\sqrt{\sum_{k=1}^{N}\tilde{u}_k{}^2}\right)\left(1-\sqrt{\sum_{k=1}^{N}\tilde{w}_k{}^2}\right). \qquad 3.6$$

Da cui i limiti inferiore $L_{\boldsymbol{u},\boldsymbol{v}}^{(N)}$ e superiore $U_{\boldsymbol{u},\boldsymbol{v}}^{(N)}$:

$$L_{\boldsymbol{u},\boldsymbol{v}}^{(N)} = F_{\boldsymbol{u},\boldsymbol{v}}^{(N)} - f(\lambda_N)\left(1-\sqrt{\sum_{k=1}^{N}\tilde{u}_k{}^2}\right)\left(1-\sqrt{\sum_{k=1}^{N}\tilde{w}_k{}^2}\right), \qquad 3.7$$

$$U_{\boldsymbol{u},\boldsymbol{v}}^{(N)} = F_{\boldsymbol{u},\boldsymbol{v}}^{(N)} + f(\lambda_N)\left(1-\sqrt{\sum_{k=1}^{N}\tilde{u}_k{}^2}\right)\left(1-\sqrt{\sum_{k=1}^{N}\tilde{w}_k{}^2}\right). \qquad 3.8$$

Se si conosce un insieme di autovalori e dei corrispondenti autovettori è dunque possibile stimare la soluzione del problema dell'equazione 2.4. Nel caso particolare in cui $\boldsymbol{u} = \boldsymbol{w}$, che quando $\boldsymbol{u} = \boldsymbol{w} = \boldsymbol{e}_i = [0, 0, \dots, 0, 1, 0, \dots, 0]$ corrisponde alla ricerca della *f-subgraph centrality* del nodo i-esimo, si ha che:

$$\left|\sum_{k=N+1}^{n} f(\lambda_k)\tilde{u}_k\tilde{w}_k\right| = \left|\sum_{k=N+1}^{n} f(\lambda_k)\tilde{u}_k{}^2\right| \le f(\lambda_N)\sum_{k=N+1}^{n}\tilde{u}_k{}^2$$

$$= f(\lambda_N)\left(1-\sum_{k=1}^{N}\tilde{u}_k{}^2\right), \qquad 3.9$$

cioè che l'approssimazione $F_{\boldsymbol{u},\boldsymbol{u}}^{(N)}$ rappresenta da sola un limite inferiore della soluzione dell'equazione 2.4.

Esistono metodi efficienti per il calcolo degli N autovalori ed autovettori più significativi data una matrice A. È inoltre dimostrato che all'aumentare del numero N di coppie autovalore-autovettore note l'approssimazione data dai limiti inferiore e superiore si affina:

$$L_{\boldsymbol{u},\boldsymbol{v}}^{(N)} - F_{\boldsymbol{u},\boldsymbol{v}}^{(N)} \leq L_{\boldsymbol{u},\boldsymbol{v}}^{(N+1)} - F_{\boldsymbol{u},\boldsymbol{v}}^{(N+1)} \leq 0, \qquad 3.10$$

$$U_{\boldsymbol{u},\boldsymbol{v}}^{(N)} - F_{\boldsymbol{u},\boldsymbol{v}}^{(N)} \geq U_{\boldsymbol{u},\boldsymbol{v}}^{(N+1)} - F_{\boldsymbol{u},\boldsymbol{v}}^{(N+1)} \geq 0. \qquad 3.11$$

La fattorizzazione spettrale parziale può essere utilizzata per l'identificazione di un sottoinsieme di nodi che contiene quelli con la massima *f-subgraph centrality* a partire da N coppie di autovalore-autovettore note. Per semplicità nel seguito della trattazione i limiti inferiore $L_{\boldsymbol{u},\boldsymbol{v}}^{(N)}$ e superiore $U_{\boldsymbol{u},\boldsymbol{v}}^{(N)}$ quando $\boldsymbol{u} = \boldsymbol{w} = \boldsymbol{e}_i = [0, 0, \dots, 0, 1, 0, \dots, 0]$ verranno denominati rispettivamente $L_{ii}^{(N)}$ e $U_{ii}^{(N)}$. Sia $\mathcal{L}_m^{(N)}$ l' m-esimo e massimo limite inferiore. Si definisce l'*index set* l'insieme:

$$S_m^{(N)} = \left\{ i : U_{ii}^{(N)} \geq \mathcal{L}_m^{(N)} \right\}, \; N = 1,2, \dots, n. \qquad 3.12$$

L'*index set* contiene dunque gli indici corrispondenti al sottoinsieme degli m nodi con la più alta *f-subgraph centrality*. In realtà esso può raggruppare più indici di quelli degli m nodi con la più alta *f-subgraph centrality*, ma sicuramente contiene questi ultimi. Quando la cardinalità dell'*index set* $\left|S_m^{(N)}\right|$ è pari a m allora esso coincide con l'insieme degli indici degli m nodi con la più alta *f-subgraph centrality*.

È importante sottolineare che i limiti trovati attraverso la fattorizzazione spettrale parziale costituiscono delle stime grossolane della soluzione dell'equazione 2.4, in tal caso per quanto riguarda la *f-subgraph centrality*. Inoltre il limite inferiore $L_{ii}^{(N)}$, corrispondente in tal caso a $F_{\boldsymbol{u},\boldsymbol{u}}^{(N)}$, converge alla soluzione dell'equazione 2.4 più velocemente del limite superiore $U_{ii}^{(N)}$ e, pertanto, costituisce un'approssimazione più accurata della media aritmetica tra i due limiti. Infine, un particolare ordine di limiti $L_{ii}^{(N)}$ i cui indici i appartengono all' *index set* non implica un analogo ordinamento delle *f-subgraph centrality* corrispondenti ai nodi con gli stessi indici.

3.1. Calcolo della fattorizzazione spettrale parziale

La computazione della fattorizzazione spettrale parziale richiede il calcolo di $f(\lambda_k)$ che, quando f è la funzione esponenziale e la rete è molto grande, può provocare overflow. Tale problema si può evitare sostituendo la matrice A con $(A - \mu I)$, con I matrice identità e μ stima di λ_{max}, autovalore massimo di A. Tipicamente, essendo $A \in \mathbb{R}^{n \times n}$ la matrice di adiacenza di una rete complessa con archi non orientati, non pesati e senza loop, il suo raggio spettrale ricade entro il valore $n - 1$, pertanto una buona scelta è prendere $\mu = n - 1$. Tuttavia, attraverso la fattorizzazione spettrale parziale è possibile stimare un insieme di N autovalori in cui sono compresi quelli dominanti. Dunque una buona scelta in tal caso è quella di prendere $\mu = \lambda_1$, con λ_1 massimo autovalore tra quelli stimati con la fattorizzazione spettrale parziale.

Un altro problema del calcolo della fattorizzazione spettrale parziale è che non è noto a priori il numero N di coppie autovalore-autovettore sufficienti a fornire dei limiti utili. Una possibile soluzione è l'impiego del metodo di Lanczos a blocchi reiterato. Tale metodo consente di trovare un insieme delle prime q coppie autovalore-autovettore dominanti. Se i limiti derivanti da tali coppie di autovalore-autovettore non sono soddisfacenti ($\left|S_m^{(q)}\right|$ è maggiore di m), allora è possibile trovare, tramite una nuova esecuzione del metodo di Lanczos a blocchi, le q coppie successive di autovalore-autovettore dominanti. Se non accade che

$\left|S_m^{(N)}\right| = m$ per $N \leq 2q$ allora si procede continuando a iterare con il metodo e calcolando, di volta in volta, q ulteriori coppie di autovalore-autovettore.

Quando la matrice di adiacenza A della rete complessa è molto grande e il numero N di coppie autovalore-autovettore necessarie al calcolo di limiti utili è anch'esso grande, ci possono essere problemi di memorizzazione delle coppie nell'esecuzione del metodi di Lanczos a blocchi reiterato. Per ovviare al problema è possibile memorizzare, invece che tutte le coppie autovalore-autovettore precedentemente trovate, solo quelle necessarie al metodo di Lanczos per il calcolo delle q coppie successive. Tuttavia in tale situazione il metodo ha complessità computazionale lievemente maggiore del caso in cui esso disponga di tutte le coppie precedentemente trovate.

Un ulteriore parametro che influisce sulla durata e sulla precisione della fattorizzazione spettrale parziale attraverso il metodo di Lanczos a blocchi reiterato è la condizione di stop. Il requisito $\left|S_m^{(N)}\right| = m$ per il numero N di coppie autovalore-autovettore da calcolare è detto *strong convergence condition* e garantisce il fatto che $S_m^{(N)}$ contenga gli m indici dei nodi con *f-subgraph centrality* maggiore. Quando tuttavia il numero di coppie richieste N è molto grande, può essere utile l'impiego della *weak convergence condition*. Tale criterio di stop prevede la terminazione dell'esecuzione del metodo di Lanczos quando i limiti inferiori $L_{ii}^{(N)}$ non aumentano significativamente all'aumentare di N. La *weak convergence condition* tuttavia può fornire *index set* $S_m^{(N)}$ che contengono più indici di nodi di quelli desiderati (m).

4. Regole di quadratura di Gauss

Le regole di quadratura del tipo di Gauss costituiscono un possibile metodo di soluzione per le equazioni del tipo 2.4. In particolare esse sono in grado di fornire un'approssimazione della soluzione migliore rispetto alla fattorizzazione spettrale parziale, ma richiedono un maggior carico computazionale.

Come visto in precedenza, quando la matrice di adiacenza A, relativa ad una rete complessa, è simmetrica, è possibile procedere alla sua fattorizzazione spettrale $A = V\Lambda V^T$, con Λ matrice diagonale contenente gli autovalori di A e V matrice composta dagli autovettori di A. L'equazione 2.4 diventa pertanto:

$$\boldsymbol{u}^T f(A)\boldsymbol{w} = \boldsymbol{u}^T f(V\Lambda V^T)\boldsymbol{w} = \sum_{k=1}^{n} f(\lambda_k)\tilde{u}_k\tilde{w}_k, \qquad 4.1$$

con $\tilde{u}_k = \boldsymbol{u}^T\boldsymbol{v}_k$ e $\tilde{w}_k = \boldsymbol{v}_k^T\boldsymbol{w}$. L'ultima somma può essere considerata un integrale di Riemann-Stieltjes:

$$\sum_{k=1}^{n} f(\lambda_k)\tilde{u}_k\tilde{w}_k = \int f(t)d\lambda(t), \qquad 4.2$$

in cui $\lambda(t)$ è una funzione di distribuzione non decrescente sull'asse reale.

4.1. Polinomi ortogonali

Sia $\mathbb{P}$ l'insieme di tutti i polinomi e $\mathbb{P}_d$ l'insieme di tutti i polinomi con grado massimo d. Dati due polinomi $u, v \in \mathbb{P}$ si definisce prodotto scalare tra i polinomi in $d\lambda(t)$:

$$\langle u, v\rangle = \int u(t)v(t)d\lambda(t). \qquad 4.3$$

Una famiglia di polinomi $p_j(t)$, con j grado del polinomio, è detta ortogonale rispetto al prodotto scalare 4.3 se:

$$\langle p_j, p_k\rangle = 0,\ j \neq k, \qquad 4.4$$

$$\langle p_j, p_k\rangle \neq 0,\ j = k. \qquad 4.5$$

Se per $j = k$ il prodotto scalare è strettamente maggiore di 0 i polinomi oltre che ortogonali sono detti monici. È possibile definire una famiglia di polinomi monici ortogonali rispetto al prodotto scalare 4.3 attraverso la formula ricorsiva:

$$\begin{cases} \pi_0(t) \equiv -1 \\ \pi_1(t) \equiv 0 \\ \pi_{j+1}(t) = (t - \delta_{j+1})\pi_j(t) - \gamma_{j+1}^2\pi_{j-1}(t),\ j = 1, 2, \dots \end{cases}, \qquad 4.6$$

dove:

$$\delta_{j+1} = \frac{\langle t\pi_j, \pi_j\rangle}{\langle \pi_j, \pi_j\rangle},\ \gamma_{j+1}^2 = \begin{cases} 0,\ j = 0 \\ \frac{\langle \pi_j, \pi_j\rangle}{\langle \pi_{j-1}, \pi_{j-1}\rangle},\ j = 1,2, \dots \end{cases}, \qquad 4.7$$

A seconda della scelta della misura indotta $d\lambda(t)$ si definiscono diverse tipologie di polinomi ortogonali (Legendre, Laguerre, Hermite, Chebyshev, etc.). Per i polinomi monici ortogonali rispetto al prodotto scalare 4.3 costruiti tramite la formula ricorsiva 4.7 vale la proprietà:

$$\langle p, \pi_n\rangle = 0,\ \forall p \in \mathbb{P}_{n-1}, \qquad 4.8$$

in quanto, essendo i $\pi_j(t)$ una base per polinomi in $d\lambda(t)$, si può esprimere $p(t)$ in funzione di $\pi_j(t)$:

$$p(t) = \sum_{j=0}^{n-1} \alpha_j\pi_j(t), \qquad 4.9$$

e dunque il prodotto scalare tra $p(t)$ e $\pi_n(t)$ sarà:

$$\langle p, \pi_n\rangle = \sum_{j=0}^{n-1} \alpha_j\langle \pi_j\pi_n\rangle = 0, \qquad 4.10$$

dal momento che j è sempre diverso da n e dunque il prodotto scalare dentro la sommatoria risulta sempre nullo.

Si dimostra inoltre che se il supporto di $d\lambda(t)$ è nell'intervallo chiuso $[a, b]$, allora gli zeri dei polinomi monici ortogonali rispetto al prodotto scalare 4.3 sono reali, semplici e compresi nell'intervallo aperto $]a, b[$.

4.2. Regole di quadratura

Si è stabilito, dalla trattazione precedente, che l'equazione 2.4 di cui si vuole trovare la soluzione, può essere espressa come un integrale in $d\lambda(t)$:

$$\boldsymbol{u}^T f(A)\boldsymbol{w} = \int f(t)d\lambda(t). \qquad 4.11$$

Le regole di quadratura sono dei metodi matematici che consentono l'approssimazione di un integrale definito attraverso la somma pesata dei valori della funzione integrata in punti specifici del dominio di integrazione:

$$\int \boldsymbol{f(t)d\lambda(t)} \approx \sum_{j=1}^{n} \boldsymbol{w_j f_j}, \qquad 4.12$$

dove $f_j = f(t_j)$ e t_j punto compreso nel dominio di integrazione.

Volendo determinare i pesi w_j per cui l'equazione 2.4 risulti esatta e considerando come $f(t)$ polinomi con il più alto grado possibile, occorre risolvere il sistema di equazioni lineari:

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ t_1 & t_2 & \cdots & t_n \\ \vdots & \vdots & \ddots & \vdots \\ t_1^{n-1} & t_2^{n-1} & \cdots & t_n^{n-1} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} = \begin{bmatrix} \int d\lambda(t) \\ \int t d\lambda(t) \\ \vdots \\ \int t^{n-1} d\lambda(t) \end{bmatrix}, \qquad 4.13$$

dove ogni riga corrisponde a un diverso polinomio $f(t) = t^j$, con j grado del polinomio stesso. La matrice così creata prende il nome di matrice di Vandermonde V_n e risulta non singolare se i nodi $t_1, t_2, \dots, t_n$ sono distinti. In tale situazione è dunque possibile trovare i pesi w_j che risolvono il sistema di equazioni lineari 2.4. Tuttavia se V_n è grande la soluzione del sistema 4.14 risulta complicato.

4.3. Regola di quadratura di Gauss

Dati $t_1, t_2, \dots, t_n$ punti distinti, la matrice

$$A = \begin{bmatrix} \pi_0(t_1) & \pi_0(t_2) & \cdots & \pi_0(t_n) \\ \pi_1(t_1) & \pi_1(t_2) & \cdots & \pi_1(t_n) \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{n-1}(t_1) & \pi_{n-1}(t_2) & \cdots & \pi_{n-1}(t_n) \end{bmatrix} \qquad 4.14$$

risulta non singolare. Se $t_1, t_2, \dots, t_n$ sono gli zeri del polinomio ortogonale $\pi_n(t)$ e $w_1, w_2, \dots, w_n$ le soluzioni del sistema

$$A \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} = \begin{bmatrix} \langle \pi_0, \pi_0 \rangle \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad 4.15$$

allora $w_j > 0$ per qualunque j e:

$$\int p(t)d\lambda(t) = \sum_{j=1}^{n} w_j p(t_j), \ \forall p \in \mathbb{P}_{2n-1}. \qquad 4.16$$

L'implicazione contraria è anch'essa valida: se l'equazione 4.17 è soddisfatta allora $t_1, t_2, \dots, t_n$ sono gli zeri del polinomio ortogonale $\pi_n(t)$ e i pesi w_j soddisfano il sistema di equazioni 4.16. Inoltre non è possibile trovare dei nodi t_j e dei pesi w_j per cui l'equazione 4.17 è soddisfatta per $\forall p \in \mathbb{P}_{2n}$.

4.3.1. Calcolo della regola di quadratura di Gauss

Il calcolo della regola di quadratura di Gauss vista precedentemente richiede in primo luogo l'individuazione degli zeri del polinomio ortogonale $\pi_n(t)$. Concettualmente l'operazione è molto semplice ma risulta difficoltoso realizzare un algoritmo stabile e generico che la implementi. In Matlab questo viene tipicamente fatto attraverso il calcolo degli autovalori della matrice compagna del polinomio (comando *roots*). Tuttavia per n elevati il tempo di elaborazione può crescere pesantemente e la stabilità dell'algoritmo diminuisce.

Relativamente alla regola di quadratura di Gauss è possibile costruire la matrice tridiagonale:

$$T_n = \begin{bmatrix} \delta_1 & \gamma_2 & 0 & \cdots & 0 \\ \gamma_2 & \delta_2 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \gamma_n \\ 0 & \cdots & 0 & \gamma_n & \delta_n \end{bmatrix} \in \mathbb{R}^{n\times n}, \qquad 4.17$$

il cui polinomio caratteristico ($\det(tI - T_n)$) coincide con il polinomio ortogonale $\pi_n(t)$. Per il calcolo degli zeri di $\pi_n(t)$ è dunque sufficiente trovare gli autovalori di T_n. Data la tridiagonalità della matrice è possibile impiegare gli algoritmi QR, i quali, per questo tipo di matrici, hanno una complessità computazionale di $O(n^2)$.

Per quanto riguarda i pesi della regola di quadratura di Gauss è possibile sfruttare la stessa matrice T_n e in particolare la corrispondente fattorizzazione spettrale:

$$T_n = U\Lambda U^T, \qquad 4.18$$

dove $\Lambda = diag(t_1, t_2, \dots, t_n)$. I pesi w_j sono dati da

$$w_j = \left(\boldsymbol{e}_1^T \boldsymbol{u}_j\right),\ j = 1, 2, \dots, n \qquad 4.19$$

nel caso in cui $\int d\lambda(t) = 1$. In caso contrario occorre anteporre una costante di scaling alle parentesi nell'equazione 4.20. È dunque possibile computare sia i nodi che i pesi della regola di quadratura di Gauss attraverso una variante dell'algoritmo QR, detta algoritmo di Golub-Welsch, con complessità computazionale $O(n^2)$. Tuttavia anche tale metodo può risultare eccessivamente lento per matrici grandi (n elevati).

Nella pratica viene utilizzato l'algoritmo di Lanczos per definire una regola di quadratura di Gauss che approssimi le soluzioni dell'equazione 2.4. Data $A \in \mathbb{R}^{n\times n}$ matrice simmetrica e $\boldsymbol{v} \in \mathbb{R}^m$ vettore unitario, attraverso l iterazioni dell'algoritmo di Lanczos è possibile ottenere una base ortogonale $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_{l+1}\}$ per il sottospazio di Krylov $K_{l+1}(A, \boldsymbol{v}) = span\{\boldsymbol{v}, A\boldsymbol{v}, \dots, A^l\boldsymbol{v}\}$.

Lo pseudocodice dell'algoritmo di Lanczos è riportato di seguito:

```
v0=0, γ1=0, v1=v;
for k=1,2,…,l
        w = Avk - γkvk-1;
        δk = vkTw;
        w = w - δkvk;
```

$\gamma_{k+1} = |\mathbf{w}|;$

$\mathbf{v}_{k+1} = \mathbf{w}/\gamma_{k+1};$

Si può notare come l'algoritmo di Lanczos in sostanza riprende la relazione ricorsiva 4.7 per la creazione dei polinomi ortogonali:

$$\boldsymbol{v}_{k+1} = A\boldsymbol{v}_k - \delta_{k+1}\boldsymbol{v}_k - \gamma_{k+1}\boldsymbol{v}_{k-1}, \qquad 4.20$$

dove i polinomi ortogonali $\pi_{j+1}(t)$ sono rappresentati da vettori ortogonali $\boldsymbol{v}_{k+1}$ e il γ_{j+1}^2 diventa γ_{k+1} in quanto la relazione ricorsiva 4.7 è relativa a polinomi monici (utilizzati maggiormente per dimostrazioni matematiche) mentre l'algoritmo di Lanczos considera vettori, e dunque polinomi, normalizzati (utilizzati tipicamente nella pratica).

L'algoritmo di Lanczos permette dunque, attraverso l iterazioni, di costruire le matrici:

$$T_l = \begin{bmatrix} \delta_1 & \gamma_2 & 0 & \cdots & 0 \\ \gamma_2 & \delta_2 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \gamma_l \\ 0 & \cdots & 0 & \gamma_l & \delta_l \end{bmatrix} \in \mathbb{R}^{l\times l},\ V_l = [\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_l] \in \mathbb{R}^{n\times l}, \qquad 4.21$$

con V_l matrice di vettori ortogonali ($V_l{}^T V_l = I$). Le matrici trovate sono tali che:

$$AV_l = V_l T_l + \gamma_{l+1}\boldsymbol{v}_{l+1}\boldsymbol{e}_l{}^T,\ \boldsymbol{e}_l{}^T = [0, 0, \dots, 0, 1]. \qquad 4.22$$

Il termine $\gamma_{l+1}\boldsymbol{v}_{l+1}\boldsymbol{e}_l{}^T$ è una matrice di rango 0 e rappresenta il termine di errore compiuto nell'approssimazione dell'equazione 2.4 tramite la regola di quadratura di Gauss.

4.4. Calcolo dei limiti attraverso le regole di quadratura

Si vuole approssimare l'equazione 2.4 attraverso una regola di quadratura di Gauss G_l calcolata tramite l iterazioni dell'algoritmo di Lanczos. Se prendiamo come esempio il calcolo della *f-subgraph centrality* ad esempio si ha:

$$\boldsymbol{v}^T f(A)\boldsymbol{v} \approx G_l f \coloneqq \sum_{i=1}^{l} f(t_i) w_i, \qquad 4.23$$

dove i t_i sono i nodi della regola di Gauss, coincidenti con gli autovalori della matrice T_l, mentre i w_i sono i pesi della regola di Gauss, corrispondenti alle prime componenti degli autovettori della matrice T_l al quadrato e moltiplicati per $\|\boldsymbol{v}\|$:

$$G_l f = \|\boldsymbol{v}\| \boldsymbol{e}_1{}^T f(T_l) \boldsymbol{e}_1. \qquad 4.24$$

Dal momento che T_l è una matrice tridiagonale non singolare si può calcolare la corrispondente fattorizzazione spettrale:

$$T_l = Q_l D_l Q_l{}^T,\ Q_l = \begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1l} \\ q_{21} & q_{22} & \cdots & q_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ q_{l1} & q_{l2} & \cdots & q_{ll} \end{bmatrix}, \qquad 4.25$$

dove D_l è la matrice diagonale contenente gli autovalori di T_l, ovvero i nodi di Gauss t_i ($D_l = diag(t_1, t_2, \dots, t_l)$), mentre Q_l ne include gli autovettori. Allora:

$$\boldsymbol{e}_1{}^T f(T_l) \boldsymbol{e}_1 = \boldsymbol{e}_1{}^T f\left(Q_l D_l Q_l{}^T\right) \boldsymbol{e}_1 = \boldsymbol{e}_1{}^T Q_l f(D_l) Q_l{}^T \boldsymbol{e}_1 = \sum_{i=1}^{l} f(t_i) q_{1i}^2$$

$$= \textstyle\sum_{i=1}^{l} f(t_i) q_{1i}^2. \qquad 4.26$$

Per il caso specifico del calcolo della *f-subgraph centrality* ($\boldsymbol{v} = \boldsymbol{w}$) e considerando la fattorizzazione spettrale della matrice A ($A = U\Lambda U^T$) l'equazione 4.27 diventa:

$$\boldsymbol{v}^T f(A) \boldsymbol{v} = \boldsymbol{v}^T f(U\Lambda U^T) \boldsymbol{v} = \boldsymbol{v}^T U f(\Lambda) U^T \boldsymbol{v} = \tilde{\boldsymbol{v}}^T f(\Lambda) \tilde{\boldsymbol{v}} = \sum_{j=1}^{m} f\left(\lambda_j\right) \tilde{\boldsymbol{v}}_j^2$$

$$= \int f(t) d\lambda(t). \qquad 4.27$$

Si dimostra che se $f(t)$ ha $2n$ derivate continue nell'inviluppo convesso del supporto $[a, b]$ della misura indotta $d\lambda(t)$, allora:

$$\int f(t) d\lambda(t) - \textstyle\sum_{i=1}^{n} w_i f(t_i) = \frac{f^{(2n)}(\xi)}{(2n)!} \langle \pi_n, \pi_n \rangle = \frac{f^{(2l)}(\xi)}{(2l)!} \int \prod_{i=1}^{l} (t - t_i)^2 \, d\lambda(t), \qquad 4.28$$

dove $a < \xi < b$. Il termine $\frac{f^{(2n)}(\xi)}{(2n)!} \langle \pi_n, \pi_n \rangle$ costituisce dunque l'errore compiuto nell'approssimazione dell'integrale attraverso la regola di Gauss. Se si conosce il segno di tale errore, si può evincere se l'approssimazione è un limite inferiore o superiore. Se, per esempio, $f(t) = e^t$ allora $f^{(2l)}(t) = e^t$. Non solo la condizione di esistenza e continuità delle $2l$ derivate di $f(t)$ nell'inviluppo convesso del supporto di $d\lambda(t)$ è soddisfatta, ma tali derivate sono anche sempre strettamente maggiori di zero. Per tale motivo:

$$\boldsymbol{v}^T f(A) \boldsymbol{v} - G_l f = \frac{f^{(2l)}(\xi)}{(2l)!} \int \prod_{i=1}^{l} (t - t_i)^2 \, d\lambda(t) > 0. \qquad 4.29$$

In altri termini la regola di quadratura di Gauss costituisce un limite inferiore per il calcolo di $\boldsymbol{v}^T f(A) \boldsymbol{v}$ quando $f(t)$ è la funzione esponenziale. Inoltre vale la relazione:

$$G_{l-1}f \leq G_l f \leq \boldsymbol{v}^T f(A)\boldsymbol{v}. \qquad 4.30$$

Se si aumentano le iterazioni l dell'algoritmo di Lanczos, si ottengono limiti inferiori sempre più accurati. In generale è necessario che $f^{(2l)}(t)$ non cambi segno nel supporto di $d\lambda(t)$ per stabilire se la regola di Gauss è un limite inferiore o superiore.

Il calcolo del limite superiore richiede una modifica della regola di quadratura di Gauss attraverso l'aggiunta di un nodo noto fissato ad un estremo del supporto di $d\lambda(t)$. Tale variante alla regola di quadratura di Gauss prende il nome di regola di quadratura di Gauss-Radau. Il problema si riconduce sempre all'approssimazione di un integrale definito attraverso una regola di quadratura di Gauss:

$$\int f(t)d\lambda(t) = \sum_{i=1}^{l} w_i f(t_i) + \varepsilon_l(f). \qquad 4.31$$

Tramite l'errore $\varepsilon_l(f)$ è possibile capire se la regola di quadratura è un limite inferiore o superiore. Si definisce in tale contesto il polinomio modale $w_l(t)$ come:

$$w_l(t) = \prod_{i=1}^{l}(t - t_i). \qquad 4.32$$

Si dimostra che, dato un intero k tale che $0 \leq k \leq l$, la regola di quadratura 4.32 risulta esatta per tutti i polinomi in $\mathbb{P}_{l-1+k}$ se e solo se: *a)* la regola è esatta per tutti i polinomi in $\mathbb{P}_{l-1}$ e *b)* il polinomio modale soddisfa l'equazione:

$$\int w_l(t)p(t)d\lambda(t) = 0, \ \forall p \in \mathbb{P}_{k-1}. \qquad 4.33$$

Assumendo il supporto di $d\lambda(t)$ in $[a, b]$, con $a > -\infty$, si consideri una regola di quadratura con un nodo $t_0 = a$ fissato a priori. Il polinomio modale corrispondente sarà:

$$w_l(t) = (t - a)\prod_{i=1}^{l}(t - t_i) = (t - a)w_l(t). \qquad 4.34$$

Per avere una regola di quadratura che sia esatta per polinomi con grado maggiore possibile allora deve valere la 4.34:

$$\int w_l(t)p(t)(t - a)d\lambda(t) = \int w_l(t)p(t)d\lambda_a(t) = 0, \ \forall p \in = 0, \ \forall p \in \mathbb{P}_{l-1}, \qquad 4.35$$

in cui $k = l$ e $d\lambda_a(t) = (t - a)d\lambda(t)$ è una nuova misura indotta non negativa se $d\lambda(t)$ è non negativa. Inoltre $w_l(t)$è un polinomio monico ortogonale nella nuova misura $d\lambda_a(t)$. La regola di quadratura a $l + 1$ nodi risultante sarà:

$$\int f(t)d\lambda(t) = w_0 f(a) + \sum_{i=1}^{l} w_i f(t_i) + \varepsilon_{l+1,a}(f), \qquad 4.36$$

dove $\varepsilon_{l+1,a}(f) = 0$ per $\forall f \in \mathbb{P}_{2l}$. Inoltre sia i nodi t_i che i pesi w_0 e w_i dipendono da a. Dalla 4.29, si ha che:

$$\varepsilon_{l+1}(f) = \frac{f^{(2l+1)}(\xi)}{(2l+1)!}\int \prod_{i=1}^{l}(t - t_i)^2 \, d\lambda_a(t) = \frac{f^{(2l+1)}(\xi)}{(2l+1)!}\int (t - a)\prod_{i=1}^{l}(t - t_i)^2 \, d\lambda(t), \qquad 4.37$$

dove $a < \xi < b$. Se si ha $f(t) = e^{-t}$ allora $f^{(2l+1)}(t) = -e^t$, sempre negativa nel supporto di $d\lambda(t)$. Pertanto, dal momento che la quantità all'interno dell'integrale è sempre positiva, l'errore di approssimazione $\varepsilon_{l+1}(f)$ risulta essere negativo: la regola di quadratura di Gauss-Radau costituisce in tal

caso un limite superiore per il calcolo di $\boldsymbol{v}^T f(A)\boldsymbol{v}$. Si noti che, scegliendo $f(t) = e^t$ e $t_0 = b$, $f^{(2l+1)}(t) = e^t$, sempre positiva nel supporto di $d\lambda(t)$, e la quantità dentro l'integrale risulta sempre negativa, per cui la regola di quadratura di Gauss-Radau costituisce ancora un limite superiore per il calcolo di $\boldsymbol{v}^T f(A)\boldsymbol{v}$ quando $f(t)$ è la funzione esponenziale.

In generale, conoscendo l'andamento delle derivate di $f(t)$ nel supporto di $d\lambda(t)$ e scegliendo opportunamente il nodo noto t_0 è possibile identificare la tipologia di limite che la regola di quadratura di Gauss-Radau rappresenta per il calcolo di $\boldsymbol{v}^T f(A)\boldsymbol{v}$.

4.5. Calcolo della regola di quadratura di Gauss-Radau

La regola di Gauss a l nodi, come visto, può essere calcolata tramite il calcolo degli autovalori e dei primi elementi degli autovettori normalizzati relativi alla matrice T_l, ottenuta attraverso l iterazioni dell'algoritmo di Lanczos. In maniera analoga si può definire una matrice $\tilde{T}_{l+1}$ come:

$$\tilde{T}_{l+1} = \begin{bmatrix} T_l & \gamma_{l+1} \\ \gamma_{l+1} & \tilde{\delta}_{l+1} \end{bmatrix} \in \mathbb{R}^{l+1\times l+1}, \qquad 4.38$$

dove γ_{l+1}viene ottenuto tramite l'algoritmo di Lanczos convenzionale, mentre $\tilde{\delta}_{l+1}$ non corrisponde a un normale δ_{l+1} ma viene calcolato in modo tale da far sì che $\tilde{T}_{l+1}$ abbia un autovalore uguale al nodo fissato t_0. Ciò non è altro che un problema di calcolo di autovalori del tipo:

$$\tilde{T}_{l+1}\boldsymbol{x} = a\boldsymbol{x}, \qquad 4.39$$

dove a è l'autovalore voluto e $\boldsymbol{x}$ il rispettivo autovettore. La soluzione del problema è data da:

$$\tilde{\delta}_{l+1} = t_0 - \gamma_{l+1}\frac{\pi_{l-1}(t_0)}{\pi_l(t_0)}. \qquad 4.40$$

5. Metodo ibrido e risultati sperimentali

Sono stati presentati due diversi metodi per la risoluzione di equazioni in forma analoga alla 2.4 ed a cui sono riconducibili diverse grandezze di interesse nell'analisi delle reti. Tali metodi, la fattorizzazione spettrale parziale e le regole di quadratura di Gauss, risultano essere particolarmente utili quando le reti con cui si ha a che fare, e dunque le relative matrici di adiacenza A, sono molto grandi. In queste situazioni infatti il calcolo diretto di $f(A)$ può essere impraticabile, sia in termini di tempo che in termini di risorse, in primo luogo di memorizzazione, necessarie. Metodi alternativi come quelli proposti possono essere sfruttati in tal senso per ottenere una buona approssimazione della soluzione cercata in tempi accettabili.

La fattorizzazione spettrale parziale consente di computare la soluzione di un'equazione come la 2.4 in tempi molto veloci, a seconda della condizione di convergenza (*strong* o *weak*) imposta. Di contro la questo metodo spesso non è in grado di fornire approssimazioni della soluzione sufficientemente accurate. Le regole di quadratura di Gauss sono invece capaci di dare una stima della soluzione molto vicina al valore

esatto. Tuttavia, se si hanno matrici A grandi, il tempo di calcolo della soluzione, seppur inferiore a quello necessario per il calcolo diretto di $f(A)$, può essere elevato.

Nel lavoro di Fenu et. al [1] è stato proposto un metodo ibrido che sfrutta le peculiarità della fattorizzazione spettrale parziale e delle regole di quadratura di Gauss per risolvere equazioni della stessa forma della 2.4. In particolare il metodo ibrido prevede dapprima l'applicazione della fattorizzazione spettrale parziale in modo da identificare un sottoinsieme di nodi della rete entro cui potrebbe essere la soluzione o le soluzioni cercate. La fattorizzazione spettrale fornisce solo una stima grossolana della soluzione dell'equazione 2.4 per i nodi identificati. La stima viene affinata, in un secondo momento, attraverso l'applicazione delle regole di quadratura di Gauss. Tale metodo agisce solo su un sottoinsieme dei nodi di A e risulta dunque meno affetto dalla grandezza della matrice di adiacenza di partenza. Il metodo ibrido sfrutta dunque i vantaggi dei due metodi: la velocità di calcolo della fattorizzazione spettrale parziale, per ridurre la dimensione del problema, e le regole di quadratura di Gauss, per ottenere una precisione elevata. Ovviamente il metodo risente in alcuni casi, come si vedrà in seguito, anche dei problemi legati ai due metodi.

5.1. Le reti di test

In questa sezione vengono descritte le reti prese in considerazione per i test sperimentali svolti. Sono state adottate delle reti reali di diversa natura e dimensione. In particolare si ha:

- *yeast* (2114 nodi, 4480 archi) – rete relativa all'interazione tra proteine nel lievito: ciascun arco rappresenta la relazione tra due proteine, corrispondenti invece ai nodi;
- *power* (4941 nodi, 13188 archi) – rete elettrica della parte occidentale degli Stati Uniti (non orientata e non pesata);
- *internet* (22963 nodi, 96872 archi) – rete che illustra la struttura di internet al 22 luglio 2006 a livello di sistemi autonomi;
- *collaboration* (40421 nodi, 351304 archi) – rete di collaborazione tra scienziati che hanno postato degli articoli in prestampa nell'archivio del sito www.arXiv.com dal primo gennaio 1995 al 31 marzo 2005;
- *facebook* (63731 nodi, 1545686 archi) – rete di amicizie tra gli utenti Facebook della città di New Orleans.

Tutte le matrici di adiacenza A delle reti prese in esame sono simmetriche di dimensione $n \times n$, con n pari al numero di nodi della rete stessa. Il numero di archi delle reti dà invece un'idea del livello di sparsità della matrice di adiacenza corrispondente. Tale livello incide sia sulla complessità computazionale degli algoritmi che sulla memoria necessaria ad eseguirli. È necessario sottolineare tuttavia che ad una matrice A con alto livello di sparsità non corrisponde una matrice $f(A)$ con un altrettanto alto livello di sparsità.

Un'altra metrica importante per la valutazione dei metodi proposti è la distribuzione e separazione degli autovalori dominanti della matrice A. Questa può infatti influire sulla velocità di convergenza e sulla correttezza del risultato ottenuto per quanto riguarda i metodi che impiegano la fattorizzazione spettrale parziale, che calcola appunto un sottoinsieme degli autovalori di A al fine di approssimare la soluzione dell'equazione 2.4. Tuttavia la distribuzione degli autovalori di A non è sufficiente a capire la velocità di convergenza del metodo o la sua accuratezza, in quanto essa può differire sostanzialmente dalla

distribuzione della misura sotto esame, dalla cui stima (limite superiore e inferiore) dipende direttamente il criterio di convergenza. Per quanto riguarda le rete *yeast*, si considerino la Figura 1, che riporta la distribuzione degli autovalori della matrice di adiacenza ad essa corrispondente, e la Figura 2, che prende in considerazione invece la misura di *f-subgraph centrality* con f funzione esponenziale matriciale. Si può notare come, mentre gli autovalori dominanti risultano tutti abbastanza vicini, per quanto riguarda la *f-subgraph centrality* vi è un nodo che riporta di gran lunga il valore più elevato rispetto agli altri. La distribuzione degli autovalori dominanti successivi al primo invece è abbastanza simile a quella dei nodi con più alta *f-subgraph centrality*.

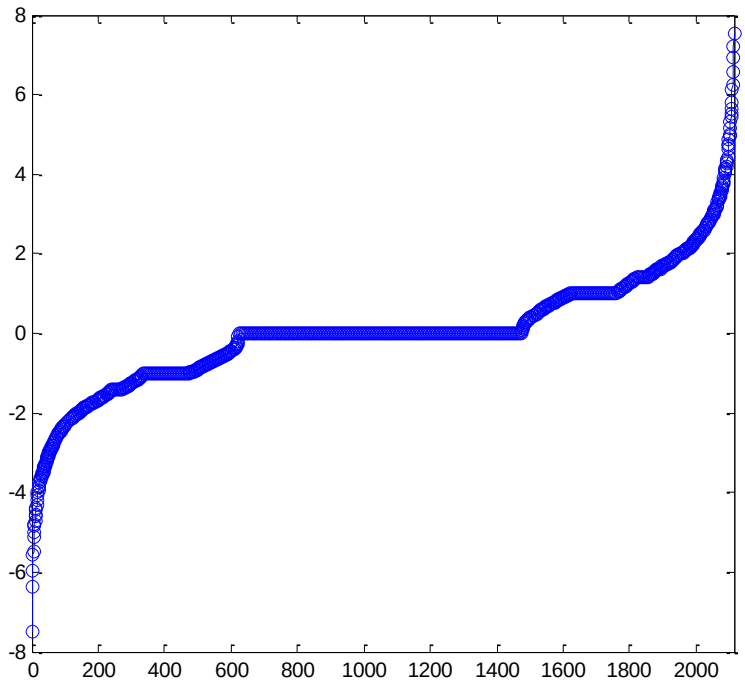

Figura 1 Distribuzione degli autovalori della matrice di adiacenza relativa alla rete *yeast*.

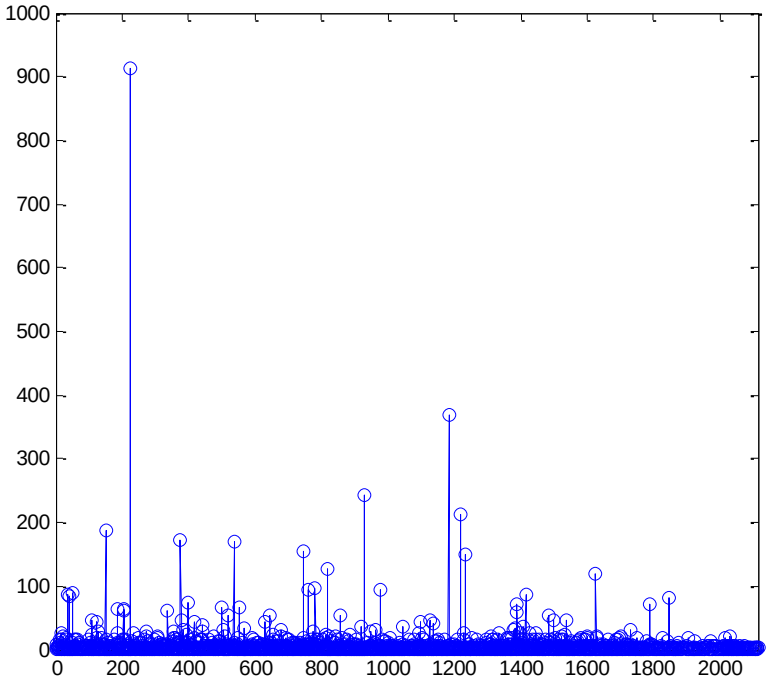

Figura 2 Distribuzione della misura di *f-subgraph centrality* ($\boldsymbol{f(A) = e^A}$) relativamente alla rete *yeast*.

Se si considerano le distribuzioni di autovalori e *f-subgraph centrality* relative alla rete *power* (rispettivamente Figura 3 e Figura 4) si può notare come, sebbene vi siano due autovalori dominanti ben distinti, per quanto riguarda la *f-subgraph centrality* un solo nodo risulta avere un valore elevato e isolato rispetto agli altri. Il secondo valore più alto di *f-subgraph centrality* è invece conteso da tre diversi nodi. Ancora una volta dunque le due distribuzioni non hanno una corrispondenza esatta.

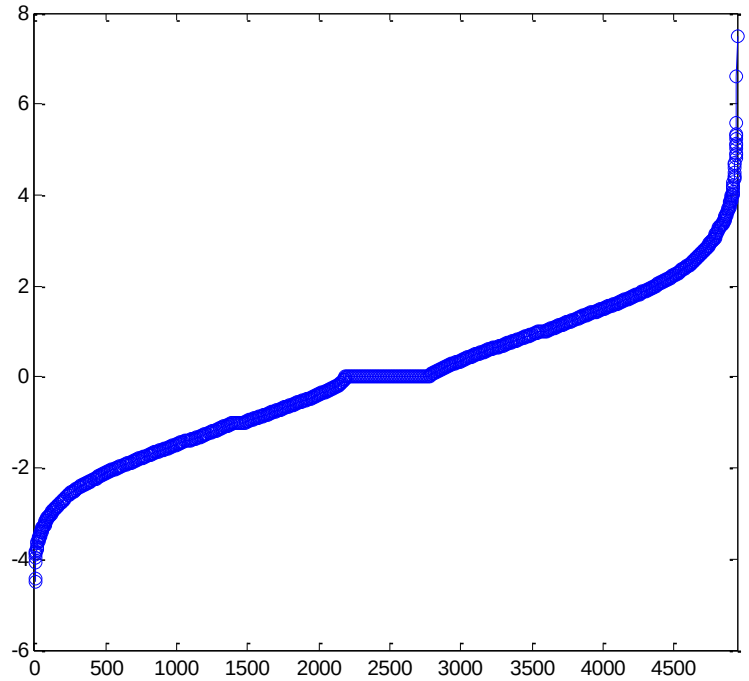

Figura 3 Distribuzione degli autovalori della matrice di adiacenza relativa alla rete *power*.

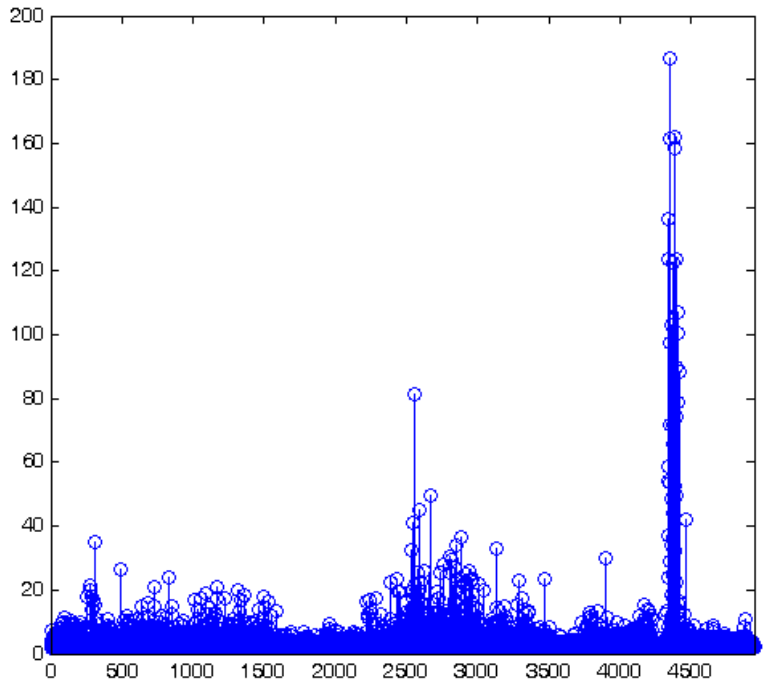

Figura 4 Distribuzione della misura di *f-subgraph centrality* ($\boldsymbol{f(A) = e^A}$) relativamente alla rete *power*.

5.2. Risultati Sperimentali

Sono stati effettuati diversi esperimenti sulle reti di test introdotte nella sezione precedente. L'ambiente di test adottato è quello fornito dal software di calcolo Matlab. Per il calcolo delle grandezze di interesse, riconducibile alla risoluzione dell'equazione 2.4, sono stati utilizzati i seguenti metodi:

- *direct* – calcolo diretto della funzione matriciale $f(A)$ attraverso i costrutti disponibili sull'ambiente di test;
- *spectral* – fattorizzazione spettrale parziale, sulla base dei concetti descritti nella sezione 3;
- *gauss* – regole di quadratura di Gauss, sulla base dei concetti descritti nella sezione 4;
- *hybrid* – metodo ibrido, proposto da Fenu et. al [1], che combina i metodi *spectral* e *gauss*.

In una prima sessione (sezione 5.2.1) di test sono stati utilizzati tali metodi per l'identificazione dei nodi più significativi di una data rete in termini di *f-subgraph centrality*. Si è analizzato il tempo di computazione e la correttezza dei metodi al variare della rete e del numero di nodi più importanti da identificare. In un secondo momento (sezione 5.2.2) si è fatto invece un confronto tra la metrica di *f-subgraph centrality* e quella del grado, entrambe relative all'importanza di un nodo nella rete, per analizzarne le principali differenze nella classificazione dei nodi della rete.

5.2.1. Calcolo della *f-subgraph centrality*

Nel tentativo di valutare i metodi proposti è stato considerato il calcolo della *f-subgraph centrality* per le reti prese in esame. In particolare è stata adottata una f corrispondente alla funzione matriciale esponenziale $f(A) = e^A$. In tal caso il metodo *direct* è costituito dal calcolo di $f(A)$ (comando *expm*), dall'estrazione dei corrispondenti elementi della diagonale (comando *diag*) e dall'ordinamento degli stessi (comando *sort*). Per quanto riguarda invece il metodo *spectral*, impiegato anche all'interno di quello *hybrid*, è stato adottato un criterio di convergenza *weak*.

Dapprima si è fatta un'analisi dei risultai ottenuti attraverso i metodi in esame al variare della rete cui la matrice di adiacenza A è riferita. Tale analisi può dare un idea del comportamento dei metodi al variare della dimensione del problema. Il numero di nodi con più alta *f-subgraph centrality* da identificare è stato fissato a 5. I risultati ottenuti sono stati riportati in Tabella 1.

Tabella 1 Tempi di esecuzione e errori commessi (in termini di numero di nodi in posizione sbagliata) dai metodi analizzati nel calcolo della *f-subgraph centrality* al variare della rete presa in esame.

rete	n	**metodo**							
		direct		***spectral***		***gauss***		***hybrid***	
		tempo	**# err**	**tempo**	**# err**	**tempo**	**# err**	**tempo**	**# err**
yeast	2114	2.7e+01	0	1.1e+00	1	3.3e+00	0	2.0e-01	0
power	4941	1.0e+02	0	5.2e-01	3	1.1e+01	0	7.9e-01	0
internet	22963	-	-	3.0e+00	0	3.3e+02	0	4.4e+00	0

collaborations	40421	-	-	9.9e+00	0	1.9e+03	0	9.1e+00	0
facebook	63731	-	-	5.7e+00	0	1.1e+04	0	1.3e+01	0

Si può subito notare dalla colonna *direct* della Tabella 1 che il calcolo diretto della *f-subgraph centrality* risulta impraticabile, a causa della durata e/o dell'eccessiva memoria richiesta, per tre delle cinque reti prese in esame. Per quanto riguarda gli altri metodi, quelli più performanti in termini di latenza sono *spectral* e *hybrid* in quanto, come ci si poteva aspettare, entrambi godono della velocità fornita dalla fattorizzazione spettrale parziale. È anche visibile il punto debole del metodo *gauss* che, al crescere della complessità del problema, richiede tempi di esecuzione molto grandi: poco più di un'ora per la rete *collaborations* e circa tre ore per *facebook*.

Se si prende in considerazione la colonna che riporta gli errori si può invece apprezzare la precisione dei metodi basati sulle regole di quadratura di Gauss (*gauss* e *hybrid*). Focalizzandosi invece su *spectral* è si può vedere la precisione limitata di tale metodo. In alcuni casi il metodo *hybrid* risente del tempo richiesto dalle regole di quadratura di Gauss per affinare la stima fornita dalla fattorizzazione spettrale parziale, risultando più lento del metodo *spectral* puro (*power*, *internet* e *facebook*). Il rallentamento rispetto a quest'ultimo metodo cresce con l'aumentare di n. Negli altri casi, *yeast* e *collaborations*, *hybrid* è il metodo più veloce in assoluto. Anche in questo caso il guadagno rispetto al metodo *spectral* diminuisce man mano che aumenta la dimensione della rete. Infine sembra che il metodo ibrido non risenta, in nessun caso, della poca precisione derivante dalla fattorizzazione spettrale parziale.

Nel tentativo di analizzare più in profondità i metodi considerati, è stato fatto variare il numero di nodi con più alta *f-subgraph centrality* da identificare. In particolare tale numero è stato fissato ai valori: 1, 5, 10, 20 e 50.

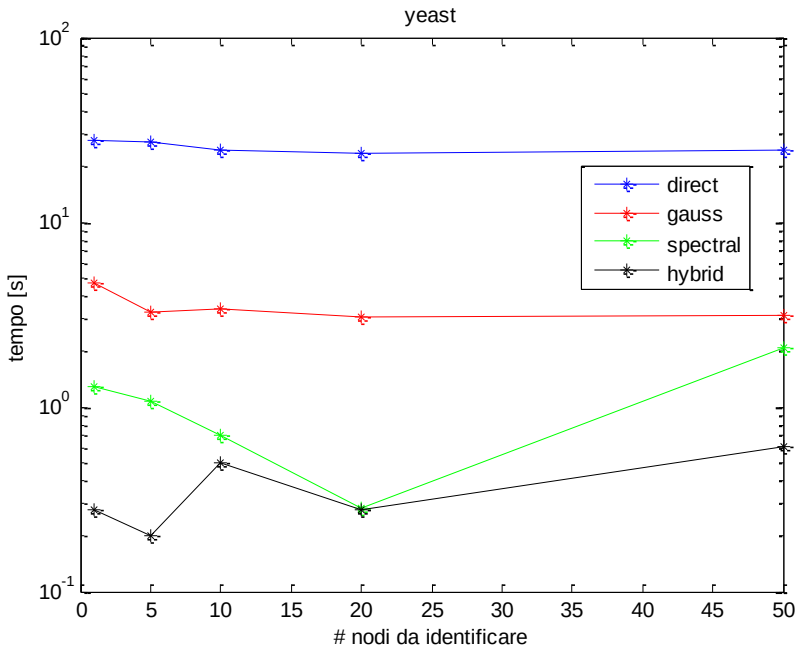

Figura 5 Tempi di esecuzione dei metodi analizzati al variare del numero di nodi da identificare per la rete *yeast*.

La Figura 5 riporta il grafico delle tempistiche di esecuzione relative alla rete *yeast*. I valori cui si riferisce il grafico sono riportati nella Tabella 2. I tempi di esecuzione hanno l'andamento che ci si aspettava dalla trattazione teorica: il metodo ibrido risulta essere quello più veloce, seguito dal metodo *spectral*, dal *gauss* e infine da quello diretto. Se si guardano gli errori di classificazione commessi, si nota subito il punto debole del metodo *spectral*, che nel caso peggiore (50 nodi da identificare) compie ben 33 errori di classificazione. Tale debolezza influisce sul metodo ibrido ma solamente nel caso di 50 nodi da identificare dove risulta esserci un errore di classificazione.

Tabella 2 Tempi di esecuzione e errori commessi (in termini di numero di nodi in posizione sbagliata) dai metodi analizzati nel calcolo della *f-subgraph centrality* per la rete *yeast* al variare del numero dei nodi da identificare.

# nodi da identificare	**metodo**							
	direct		***spectral***		***gauss***		***hybrid***	
	tempo	**# err**	**tempo**	**# err**	**tempo**	**# err**	**tempo**	**# err**
1	2.8e+01	0	1.3e+00	0	4.7e+00	0	2.8e-01	0
5	2.7e+01	0	1.1e+00	1	3.3e+00	0	2.0e-01	0
10	2.4e+01	0	7.0e-01	0	3.4e+00	0	5.0e-01	0
20	2.4e+01	0	2.9e-01	6	3.1e+00	0	2.7e-01	0
50	2.4e+01	0	2.1e+00	33	3.1e+00	0	6.1e-01	1

La Figura 6 e la Tabella 3 riportano invece le latenze e gli errori dei metodi in esame per quanto riguarda la rete *power*. Si può notare come in tal caso il metodo *spectral* risulta più veloce di quello *hybrid* in tre dei cinque valori di numero di nodi da identificare testati. Il metodo basato sulle regole di quadratura di Gauss, utilizzato nel metodo ibrido per affinare la stima delle soluzioni, sta rallentando significativamente l'esecuzione.

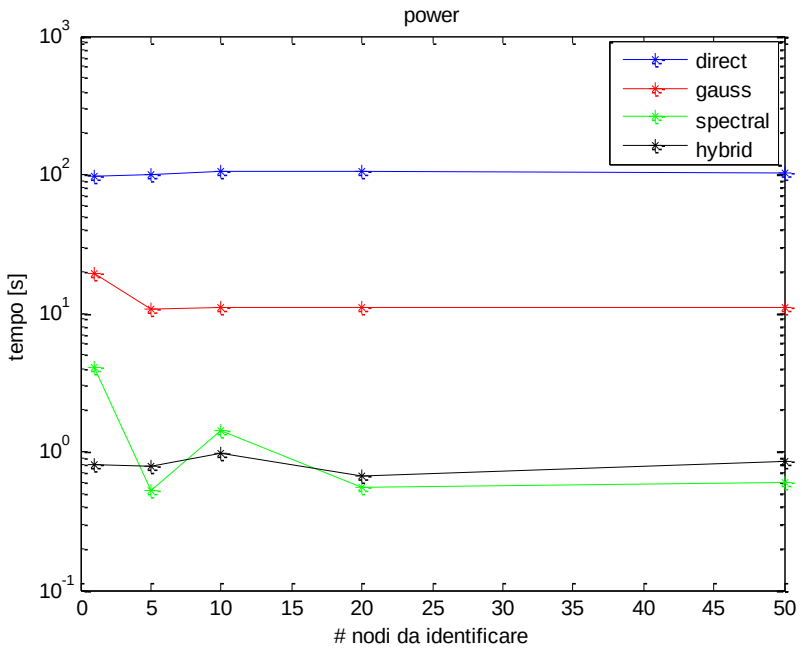

Figura 6 Tempi di esecuzione dei metodi analizzati al variare del numero di nodi da identificare per la rete *power*.

Ancora una volta il metodo *spectral* si rivela impreciso, dando luogo ad errori in quasi tutte le configurazioni testate. Per la situazione in cui vi sono 50 nodi da identificare anche il metodo hybrid risente della poca accuratezza della fattorizzazione spettale parziale.

Tabella 3 Tempi di esecuzione e errori commessi (in termini di numero di nodi in posizione sbagliata) dai metodi analizzati nel calcolo della *f-subgraph centrality* per la rete *power* al variare del numero dei nodi da identificare.

# nodi da identificare	**metodo**							
	direct		***spectral***		***gauss***		***hybrid***	
	tempo	**# err**	**tempo**	**# err**	**tempo**	**# err**	**tempo**	**# err**
1	9.8e+01	0	4.1e+00	0	1.9e+01	0	8.2e-01	0
5	1.0e+02	0	5.2e-01	3	1.1e+01	0	7.9e-01	0
10	1.1e+02	0	1.4e+00	6	1.1e+01	0	9.8e-01	0
20	1.0e+02	0	5.6e-01	14	1.1e+01	0	6.6e-01	0
50	1.0e+2	0	5.9e-01	33	1.1e+01	0	8.5e-01	8

La Figura 7 e la Tabella 4 si riferiscono alle misure effettuate sulla rete *internet*. Tale rete costituisce il primo caso in cui il metodo diretto non è applicabile a causa degli eccessivi tempi e risorse richiesti. Il metodo ibrido e quello *spectral* hanno anche in tal caso tempi di esecuzione molto simili e di gran lunga inferiori a quelli del metodo *gauss*. Il metodo *spectral* risulta più performante in quattro su cinque configurazioni di test.

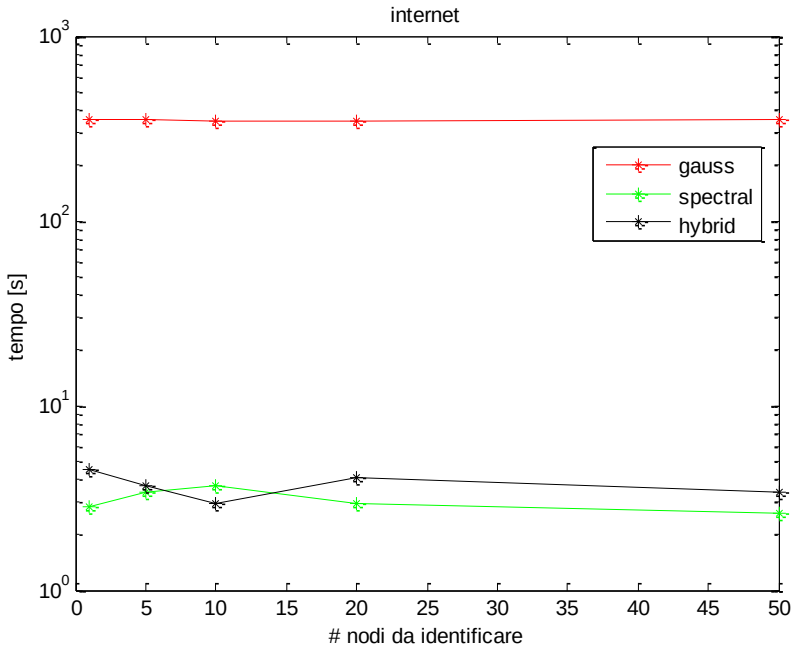

Figura 7 Tempi di esecuzione dei metodi analizzati al variare del numero di nodi da identificare per la rete *internet*.

Per questa particolare rete e per i numeri di nodi da identificare fissati il metodo *spectral* non commette errori di classificazione. Pertanto in tale contesto, date anche le migliori prestazioni in termini di latenza, è da considerarsi preferibile rispetto al metodo ibrido.

Tabella 4 Tempi di esecuzione e errori commessi (in termini di numero di nodi in posizione sbagliata) dai metodi analizzati nel calcolo della *f-subgraph centrality* per la rete *internet* al variare del numero dei nodi da identificare.

# nodi da identificare	**metodo**							
	direct		***spectral***		***gauss***		***hybrid***	
	tempo	**# err**	**tempo**	**# err**	**tempo**	**# err**	**tempo**	**# err**
1	-	-	2.9e+00	0	3.5e+02	0	4.5e+00	0
5	-	-	3.4e+00	0	3.5e+02	0	3.7e+00	0
10	-	-	3.7e+00	0	3.5e+02	0	3.0e+00	0
20	-	-	3.0e+00	0	3.5e+02	0	4.1e+00	0
50	-	-	5.1e+00	0	4.0e+2	0	4.0e+00	0

La Figura 8 e la Tabella 5 mostrano le misure effettuate sulla rete *collaboration*. Ancora una volta *spectral* e *hybrid* hanno un comportamento simile, ma in tal caso è *hybrid* ad essere più veloce nella maggior parte dei casi (tre su due) testati.

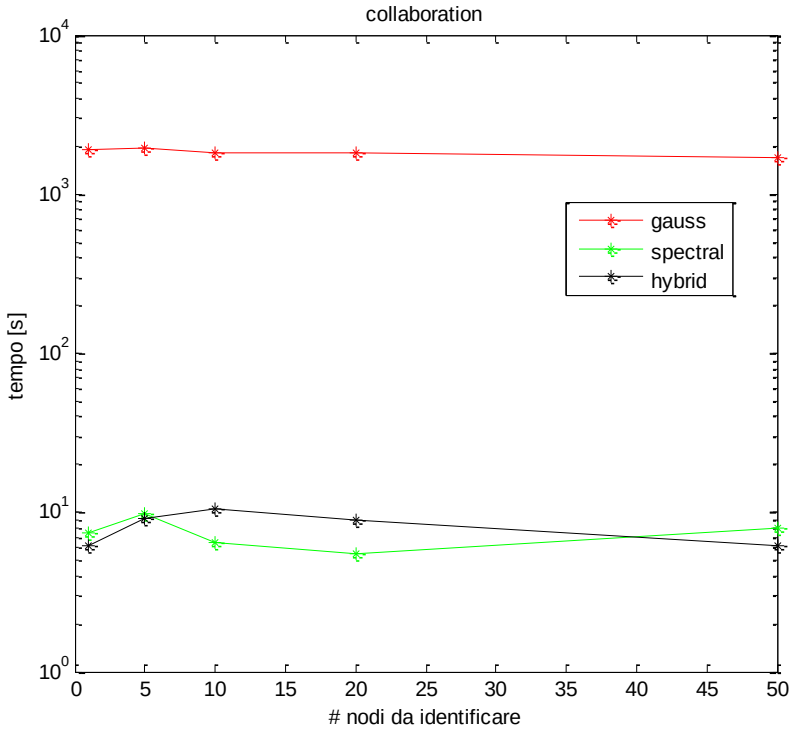

Figura 8 Tempi di esecuzione dei metodi analizzati al variare del numero di nodi da identificare per la rete *collaboration*.

Il metodo *spectral* anche per la rete *collaborations* non dà luogo ad errori di classificazione dei nodi.

Tabella 5 Tempi di esecuzione e errori commessi (in termini di numero di nodi in posizione sbagliata) dai metodi analizzati nel calcolo della *f-subgraph centrality* per la rete *collaboration* al variare del numero dei nodi da identificare.

# nodi da identificare	Metodo							
	direct		***spectral***		***gauss***		***hybrid***	
	tempo	**# err**	**tempo**	**# err**	**tempo**	**# err**	**tempo**	**# err**
1	-	-	7.4e+00	0	1.9e+03	0	6.2e+00	0
5	-	-	9.9e+00	0	1.9e+03	0	9.1e+00	0
10	-	-	6.4e+00	0	1.8e+03	0	1.0e+01	0
20	-	-	5.5e+00	0	1.8e+03	0	9.0e+00	0
50	-	-	8.0e+00	0	1.7e+3	0	6.2e+00	0

Infine la Figura 9 e la Tabella 6 mostrano i risultati ottenuti dai metodi in questione sulla rete *facebook*. Si può notare come in tal caso il metodo *spectral* risulti sempre più veloce di quello ibrido. In generale, per le reti testate, questa situazione si verifica tanto più quanto le dimensioni della rete crescono, a prescindere dal numero di nodi da identificare richiesto.

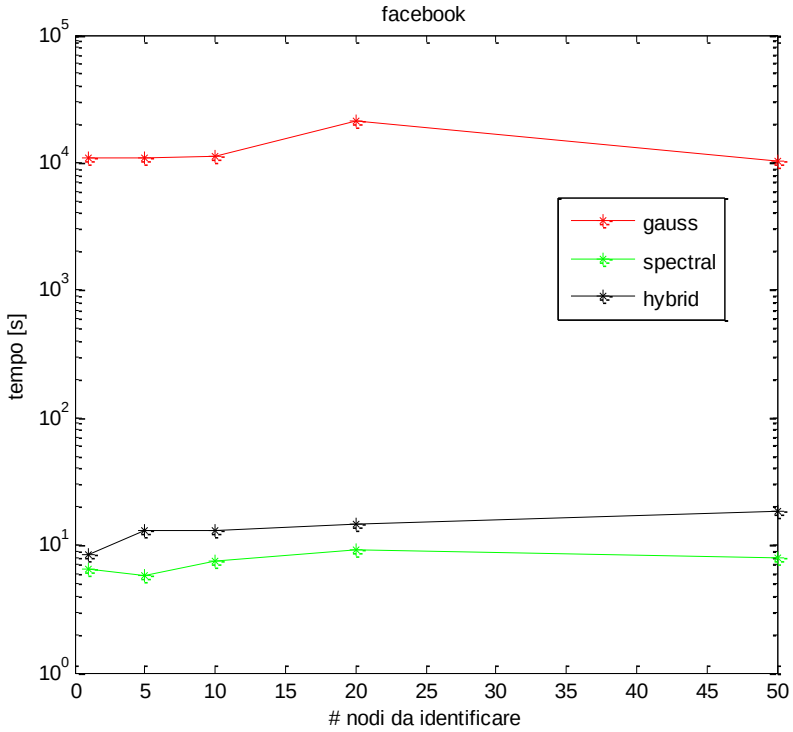

Figura 9 Tempi di esecuzione dei metodi analizzati al variare del numero di nodi da identificare per la rete *facebook*.

Si può notare inoltre come il metodo *spectral* non commetta alcun errore nella classificazione dei nodi. Questo aspetto sembra ancora essere legato alle dimensioni della rete in esame, visto che solo per le due reti più piccole (*yeast* e *power*) si hanno degli errori di classificazione da parte dei metodi che coinvolgono la fattorizzazione spettrale parziale. Tuttavia esso dipende in realtà dalla separazione dei valori di *f-subgraph centrality* relativi ai nodi. All'aumentare della dimensione di una rete aumenta anche il massimo valore di *f-subgraph centrality* potenziale che può essere raggiunto, anche se di contro aumenta anche il numero di nodi della rete che devono essere classificati. La dimensione della rete può dunque favorire una maggior accuratezza dei metodi basati sulla fattorizzazione spettrale parziale, ma essa dipende soprattutto dalla topologia della rete. Per le reti più grandi studiate la topologia è tale da avere i 50 nodi più importanti in termini di *f-subgraph centrality* con dei valori di quest'ultima metrica ben separati tra loro.

Tabella 6 Tempi di esecuzione e errori commessi (in termini di numero di nodi in posizione sbagliata) dai metodi analizzati nel calcolo della *f-subgraph centrality* per la rete *facebook* al variare del numero dei nodi da identificare.

# nodi da identificare	**metodo**							
	direct		***spectral***		***gauss***		***hybrid***	
	tempo	**# err**	**tempo**	**# err**	**tempo**	**# err**	**tempo**	**# err**
1	-	-	6.5e+00	0	1.1e+04	0	8.5e+00	0
5	-	-	5.7e+00	0	1.1e+04	0	1.3e+01	0
10	-	-	7.4e+00	0	1.1e+04	0	1.3e+01	0
20	-	-	9.2e+00	0	2.1e+04	0	1.5e+01	0

50	-	-	8.1e+00	0	1.0e+04	0	1.9e+01	0

In generale dai test effettuati non si evince una particolare correlazione tra numero di nodi richiesti e tempi di esecuzione. La variabilità delle latenze al crescere del numero di nodi sembra essere aleatoria e probabilmente legata alle caratteristiche della rete e/o alle condizioni di carico computazionale della macchina su cui sono stati eseguiti i test. In generale si può notare un andamento dei tempi di esecuzione per il metodo *gauss* che, a parte per il caso dei 20 nodi della rete *facebook*, per una data rete rimane pressoché costante a prescindere dal numero di nodi da identificare. Il metodo *spectral* e di conseguenza anche quello ibrido hanno invece andamenti più irregolari, probabilmente dipendenti dal criterio di convergenza adottato e dalla separazione dei valori di *f-subgraph centrality* dei nodi. Sempre per i metodi basati sulla fattorizzazione spettrale parziale, per quanto riguarda la correttezza, ove il metodo presenti errori, questi aumentano all'aumentare del numero di nodi da identificare.

5.2.2. Grado e *f-subgraph centrality*

Nella sezione 2 sono state introdotte diverse metriche che vengono comunemente adottate nell'analisi delle reti. In particolare il grado (o *degree*) e la *f-subgraph centrality* sono due misure atte entrambe a quantificare l'importanza di un certo nodo nella rete. In termini di complessità computazionale il *degree* risulta essere molto più semplice rispetto alla *f-subgraph centrality*. Tuttavia il *degree* classifica un nodo considerando solamente il numero di archi cui esso è connesso, senza analizzare quanto tali archi consentano al nodo di raggiungere effettivamente gli altri nodi della rete. In altre parole il *degree* non considera la lunghezza dei cammini che partono/arrivano dal/al nodo in questione. La *f-subgraph centrality* invece tiene conto anche della lunghezza dei cammini e costituisce, in teoria, una misura più completa e sofisticata del *degree*.

In questa sezione, grazie alla vastità del numero di reti di test a disposizione e alle alte prestazioni ottenibili con i metodi presentati, si vogliono analizzare le differenze tra le due metriche utilizzate nell'analisi delle reti. In particolare verrà confrontata la classificazione dei 5 nodi più importanti delle reti sotto esame in base al *degree* con quella ottenuta invece attraverso la *f-subgraph centrality*. Mentre la prima metrica è facilmente calcolabile in maniera diretta, per la seconda è stato utilizzato il metodo ibrido.

Nella Tabella 7 viene riportata la classificazione ottenuta tramite le due metriche per la rete *yeast*. Si può notare come, per questo particolare caso, gli insiemi di nodi più importanti ottenuti differiscono per un solo nodo. Se si considera anche la posizione dei nodi in termini di importanza, le differenze salgono a tre. Inoltre le due metriche classificano come primi due nodi per importanza gli stessi nodi. Per la rete *yeast* dunque il grado sembra essere una metrica simile alla *f-subgraph centrality*. In altre parole, nella rete sotto esame, se un nodo ha molti cammini di lunghezza unitaria vi è molta probabilità che lo stesso nodo abbia molti cammini di lunghezza maggiore di uno o che sia connesso a nodi anch'essi importanti. Questo è vero almeno per alcuni tra i nodi più importanti della rete stessa.

Tabella 7 Classificazione dei primi 5 nodi più importanti in termini di *degree* e di *f-subgraph centrality* per la rete *yeast* (d1=numero di ID dei nodi differenti per il solo valore, d2=numero di ID dei nodi differenti per valore e posizione).

ordine	***degree***	***f-subgraph centrality***	**d1**	**d2**

importanza	ID nodo	valore	ID nodo	valore		
1	224	5.60e+01	224	9.12e+02		
2	1183	3.80e+01	1183	3.69e+02		
3	150	3.00e+01	929	2.43e+02	1	3
4	819	2.90e+01	1219	2.13e+02		
5	1219	2.90e+01	150	1.87e+02		

La Tabella 8 mostra la classificazione ottenuta per la rete *power*. In tal caso vi è un solo nodo comune tra gli insiemi trovati con le due metriche. Oltretutto, il nodo comune non è classificato nella stessa posizione (risulta quinto in termini di *degree* e primo in termini di *f-subgraph centrality*). Diversamente da *yeast* i nodi della rete *power* che hanno molti cammini di lunghezza unitaria hanno meno probabilità di avere cammini lunghi, ad eccezione del nodo con ID 4346. Un'altra particolarità di tale rete è l'uguaglianza e il basso valore del grado (14) di tre sui cinque nodi classificati: nella rete *power* molti nodi potrebbero avere stesso valore di grado (4939 nodi hanno grado compreso tra 0 e 14). Tale rete rappresenta dunque una situazione in cui l'utilizzo del grado per identificare i nodi più importanti non è una buona scelta.

Tabella 8 Classificazione dei primi 5 nodi più importanti in termini di *degree* e di *f-subgraph centrality* per la rete *power* (d1=numero di ID dei nodi differenti per il solo valore, d2=numero di ID dei nodi differenti per valore e posizione).

ordine	*degree*		*f-subgraph centrality*		d1	d2
importanza	ID nodo	valore	ID nodo	valore		
1	2554	1.90e+01	4346	1.87e+02		
2	4459	1.80e+01	4382	1.62e+02		
3	832	1.40e+01	4353	1.61e+02	4	5
4	3469	1.40e+01	4385	1.58e+02		
5	4346	1.40e+01	4337	1.36e+02		

La classificazione relativa alla rete *internet* è illustrata nella Tabella 9. In tal caso le metriche portano ad una classificazione simile. Si registra un valore di *degree* massimo dell'ordine delle migliaia. Anche in questa situazione dunque gran parte delle migliaia di cammini passanti per i nodi più importanti in termini di grado risultano essere cammini lunghi o, in alternativa, vi sono pochi nodi coinvolti in un così alto numero di cammini unitari.

Tabella 9 Classificazione dei primi 5 nodi più importanti in termini di *degree* e di *f-subgraph centrality* per la rete *internet* (d1=numero di ID dei nodi differenti per il solo valore, d2=numero di ID dei nodi differenti per valore e posizione).

ordine	*degree*	*f-subgraph centrality*	d1	d2

importanza	ID nodo	Valore	ID nodo	Valore		
1	4	2.39e+03	4	7.53e+29		
2	3	2.02e+03	23	4.90e+29		
3	15	1.71e+03	3	4.82e+29	1	4
4	23	1.30e+03	15	4.66e+29		
5	59	1.24e+03	27	3.96e+29		

I dati ottenuti tramite le due metriche in esame per la rete *collaboration* sono riportati in Tabella 10. In tal caso si hanno ancora valori molto alti di *f-subgraph centrality* e le differenze di classificazione rimangono limitate. I due nodi più importanti sono gli stessi per entrambe le metriche e il valore di questi ultimi rimane significativamente maggiore rispetto agli altri nodi classificati.

Tabella 10 Classificazione dei primi 5 nodi più importanti in termini di *degree* e di *f-subgraph centrality* per la rete *collaboration* (d1=numero di ID dei nodi differenti per il solo valore, d2=numero di ID dei nodi differenti per valore e posizione).

ordine	*degree*		*f-subgraph centrality*		d1	d2
importanza	ID nodo	Valore	ID nodo	Valore		
1	1887	2.78e+02	1887	7.84e+20		
2	1886	2.72e+02	1886	7.04e+20		
3	680	2.46e+02	4599	4.06e+20	2	3
4	4853	2.29e+02	2380	3.61e+20		
5	769	2.22e+02	680	3.58e+20		

L'ultima rete presa in esame è *facebook*. In Tabella 11 viene mostrata la classificazione ottenuta tramite le metriche considerate. In tal caso, come per le altre reti con numero di nodi superiore alle decine di migliaia, i valori di grado e di *f-subgraph centrality* risultano essere molto elevati. Tuttavia ora vi è un unico nodo comune, anche se non con la stessa posizione tra le due classificazioni (ID 2322). La rete *facebook* si presenta come una rete in cui vi sono nodi con cammini lunghi, o comunque connessi con altri nodi importanti, che non coincidono con i nodi che hanno più cammini unitari in assoluto.

Tabella 11 Classificazione dei primi 5 nodi più importanti in termini di *degree* e di *f-subgraph centrality* per la rete *facebook* (d1=numero di ID dei nodi differenti per il solo valore, d2=numero di ID dei nodi differenti per valore e posizione).

ordine	*degree*		*f-subgraph centrality*		d1	d2
importanza	ID nodo	Valore	ID nodo	Valore		
1	2332	1.10e+03	9904	4.93e+55	4	5

2	471	9.32e+02	2322	2.67e+55		
3	554	9.17e+02	5170	2.26e+55		
4	2322	7.97e+02	5175	2.10e+55		
5	451	7.67e+02	2362	2.04e+55		

La classificazione dell'importanza dei nodi attraverso il *degree* e la *f-subgraph centrality* dipende fortemente dalla struttura della rete sotto esame. Dalla breve analisi effettuata si può evincere che tali misure sono in realtà due misure complementari: il grado dà un'idea di quali sono i nodi che hanno più cammini di lunghezza unitaria, ovvero che comunicano direttamente col numero maggiore di altri nodi; la *f-subgraph centrality* dice invece quali sono i nodi con più cammini in generale, dando maggior peso a quelli corti. Se prese assieme le due misure possono dare un'idea della topologia della rete: se le classificazioni coincidono significa che i nodi che hanno più cammini di lunghezza unitaria sono quelli che hanno anche i cammini più lunghi in generale o che comunque sono connessi a nodi anch'essi ben connessi con il resto della rete. Se le classificazioni non coincidono significa che esistono nella rete dei nodi che hanno molti cammini di lunghezza unitaria, ma vi sono nodi che nonostante abbiano meno connessioni con gli altri nodi riescono ad avere cammini di lunghezze grandi o a raggiungere facilmente altri nodi ben connessi con il resto della rete.

6. Conclusioni

Oggigiorno le reti vengono utilizzate, talvolta anche come modello, in svariati campi applicativi. La complessità e la vastità di alcuni di questi campi si rispecchiano spesso sulle reti adottate. L'analisi di tali reti diventa dunque molto complessa, specie se si adottano le metodologie tradizionali per il calcolo delle grandezze tipiche relative alle reti stesse. Trattare reti e matrici di adiacenza molto grandi è infatti problematico per via delle tempistiche necessarie ai metodi utilizzati durante l'analisi e per la grande richiesta di risorse di memorizzazione dei dati necessari. Per tale motivo sono stati presentati due metodi alternativi a quelli tradizionali, che prevedono il calcolo diretto delle grandezze in questione. I metodi proposti sono la fattorizzazione spettrale parziale e le regole di quadratura di Gauss. Entrambi forniscono un'approssimazione della misura cercata attraverso una coppia di limiti inferiore e superiore della stessa. Esse presentano dei pro e dei contro: la fattorizzazione spettrale parziale risulta essere molto veloce ma è spesso poco precisa, mentre le regole di quadratura di Gauss sono molto accurate ma, seppur più veloci del calcolo diretto, richiedono tempi lunghi in presenza di reti molto grandi. Recenti studi hanno proposto un nuovo metodo ibrido che cerca di sfruttare i punti di forza della fattorizzazione spettrale parziale e delle regole di quadratura di Gauss nel calcolo di grandezze caratteristiche relative a reti grandi.

Sono stati testati il metodo classico, quelli alternativi e quello ibrido su un insieme di test comprendente cinque reti reali. Il campione analizzato differisce, oltre che per natura delle reti, anche per la dimensione delle stesse. Ciò ha permesso di analizzare le prestazioni dei vari metodi sotto esame al variare della scala del problema. In particolare gli esperimenti condotti si sono concentrati sul calcolo della *f-subgraph centrality* delle reti, metrica che quantifica l'importanza di un certo nodo all'interno di una rete. È stata

utilizzata tale metrica per identificare l'insieme dei nodi più importanti della rete. Anche la dimensione di tale insieme di nodi è stata fatta variare durante i test. Dai risultati ottenuti sono state in primo luogo verificate le problematiche del metodo classico e di quelli alternativi. Per quanto riguarda il metodo ibrido se ne è validata l'efficacia in gran parte delle situazioni testate. È stato notato un peggioramento delle prestazioni di tale metodo all'aumentare della dimensione della rete e del numero di nodi più importanti da identificare. In questi casi il metodo ibrido risente probabilmente dei punti deboli dei metodi alternativi che lo compongono: il rallentamento dell'esecuzione, che si accentua con il crescere della dimensione della rete, e il peggioramento della precisione, che può verificarsi quando si deve identificare un numero elevato di nodi importanti.

Infine è stata analizzata la differenza tra la metrica studiata, la *f-subgraph centrality*, ed una delle metriche più semplici dell'analisi delle reti, ovvero il grado (o *degree*). Entrambe le metriche valutano l'importanza di un certo nodo della rete ma la *f-subgraph centrality* costituisce una misura più complessa e sofisticata del grado. Attraverso alcuni esperimenti volti a classificare i nodi più importanti di una rete tramite le due metriche, è emerso che si può avere una classificazione simile con *f-subgraph centrality* e con il grado a seconda della rete sotto esame. Più in generale si è trovato che le misure non sono alternative ma complementari, in quanto analizzano due diversi aspetti della rete, e che prese insieme possono fornire informazioni più complete sulla rete stessa.

Bibliografia

[1] D. M. L. R. a. G. R. C. Fenu, «Network Analysis via Partial Spectral Factorization and Gauss Quadrature,» *Methods and Algorithms for Scientific Computing,* vol. 35, n. 4, pp. 2046-2068, 2013.