Studio numerico di metriche per reti complesse

Caterina Fenu, Giuseppe Rodriguez

Dipartimento di Matematica e Informatica, Università di Cagliari
viale Merello 92, 09123 Cagliari

in collab. con Michele Marchesi e Roberto Tonelli
(Dip. Ingegneria Elettrica e Elettronica, Univ. degli Studi di Cagliari)

Due giorni di Algebra Lineare Numerica
Genova, 16-17 Febbraio 2012

Grafi

Un grafo è formalmente una coppia di insiemi $G = (V, E)$ dove:

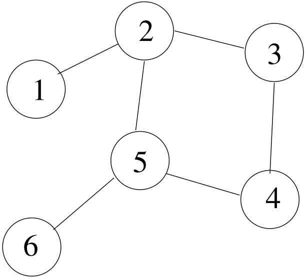
- V è un insieme finito numerabile che rappresentiamo con l'insieme $\{1, \ldots, n\}$ i cui elementi vengono chiamati nodi;
- E è un insieme di coppie $\{(i, j) \in V \times V\}$ i cui elementi vengono chiamati archi.

Se $(i, j) \in E$ allora diremo che il nodo i e il nodo j sono adiacenti.

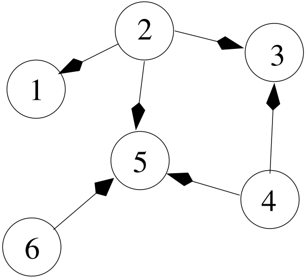
Se l'insieme E è costituito da coppie ordinate diremo che il grafo è orientato, non orientato altrimenti.

Esempi

Esempi di grafi:



Grafo non orientato



Grafo orientato

Nodi connessi e grafi connessi

Un cammino è una sequenza di archi $\{(i_1, i_2), \ldots, (i_{n-1}, i_n)\}$.
Se il grafo è orientato si parla di cammino orientato. Un cammino orientato che inizia e termina con lo stesso nodo si chiama ciclo.

Grafo non orientato: due nodi i e j si dicono connessi se esiste un cammino che porta da i a j o viceversa. Se ogni coppia di nodi è connessa allora il grafo è detto connesso.

Grafo orientato: due nodi i e j si dicono fortemente connessi se esiste un cammino orientato che porta da i a j e se questo cammino orientato esiste per ogni coppia di nodi allora il grafo è detto fortemente connesso.
Due nodi i e j si dicono invece debolmente connessi se esiste un cammino che porta da i a j nel quale ogni arco può essere percorso nel suo verso o nel verso opposto. Se ogni coppia di nodi è debolmente connessa allora il grafo è detto debolmente connesso.

Reti complesse

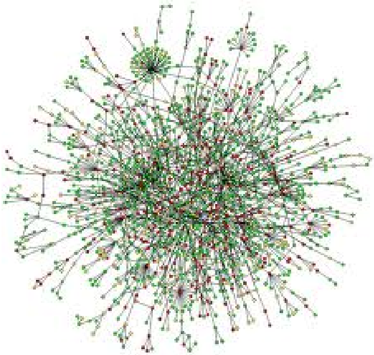
Cos'è una rete complessa?

Non esiste una definizione precisa. Una possibile scelta è quella di definire una rete complessa come un grafo in cui è presente almeno una delle seguenti caratteristiche:

- Scale-free (power law degree distribution)
- Small world
- Highly clustered.

Esempi

Esempi di reti complesse:

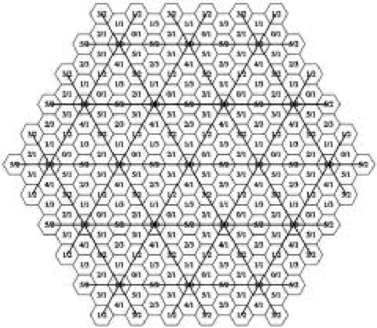


Interazioni tra proteine

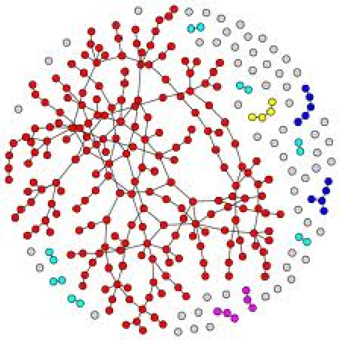
Interazioni tra neuroni

Esempi

Esempi di grafi che non sono reti complesse:



Reticolo regolare



Rete completamente random

Matrice di incidenza

Ad ogni rete complessa può essere associata una matrice di incidenza definita nel modo seguente:

$$A_{ij} = \begin{cases} w & \text{se l'arco tra il nodo i e il nodo j esiste e ha peso } w \\ 0 & \text{se il nodo i e il nodo j non sono connessi} \end{cases}$$

Lo studio delle proprietà di una rete può essere affrontato analizzando la sua matrice di incidenza che ha, infatti, caratteristiche diverse a seconda del grafo considerato.

In genere le matrici di incidenza non sono presentano particolari strutture ma un elemento comune a tutte è la sparsità.

Metriche

Con il termine metrica denotiamo un indicatore numerico che rappresenti una proprietà della rete. Dal punto di vista numerico è interessante implementare degli algoritmi efficienti che ne consentano il calcolo a partire dalla matrice di incidenza. Due questioni rilevanti sono legate alle reti complesse:

- determinare i nodi più importanti;
- determinare quanto "è facile" passare da un nodo all'altro.

Grado di un nodo

Il grado di un nodo è dato dal numero di nodi adiacenti ad esso.

Se la **Rete non è orientata** allora

$$\deg(i) = \sum_{\substack{j=1 \\ j\neq i}}^{n} A(i,j) \qquad \text{oppure} \qquad \deg(i) = \sum_{\substack{j=1 \\ j\neq i}}^{n} A(j,i)$$

Se la **Rete è orientata** per ogni nodo possiamo considerare gli archi in uscita oppure gli archi in entrata:

$$\text{deg\_out}(i) = \sum_{\substack{j=1 \\ j\neq i}}^{n} A(i,j) \qquad \text{e} \qquad \text{deg\_in}(i) = \sum_{\substack{j=1 \\ j\neq i}}^{n} A(j,i)$$

Distanza

La distanza tra due nodi è il cammino più breve che li collega (geodetica). Se due nodi non sono connessi si pone per convenzione una distanza pari a infinito.

Per effettuare il calcolo, bisogna prendere in considerazione le potenze della matrice di incidenza, infatti

$$[A^k]_{ij} = \text{n}^\circ \text{ di cammini di lunghezza k che collegano i e j}$$

Quindi:

$$\text{dist}(i,j) = \min\{k : (A^k)_{ij} \neq 0\} \qquad i \neq j$$

Subgraph centrality

La subgraph centrality indica l'importanza di un nodo in base al numero di sottografi a cui partecipa. Per fare questo si sommano i cammini chiusi quel nodo penalizzando quelli di maggiore lunghezza. In formule:

$$SC(i) = \sum_{k=0}^{\infty} \frac{[A^k]_{ii}}{k!} = [e^A]_{ii}$$

Indice di Hub e di Authority

Un nodo è un hub se punta a nodi considerati importanti e un nodo è un authority se è uno di questi nodi importanti.

Si ha quindi una definizione ricorsiva:

Un nodo ha un alto indice di hub se punta a nodi che hanno un alto indice di authority e un nodo ha un alto indice di authority se viene puntato da nodi che hanno un alto indice di hub.

Algoritmo HITS

Questo algoritmo si basa su un metodo iterativo. Ad ogni nodo i vengono assegnati due pesi non negativi: un peso di authority x_i e un peso di hub y_i a cui viene assegnato un valore iniziale arbitrario diverso da zero. I valori sono incrementati poi nel seguente modo:

$$x_i^{(k)} = \sum_{j:(j,i)\in E} y_j^{(k-1)} \qquad \text{e} \qquad y_i^{(k)} = \sum_{j:(i,j)\in E} x_j^{(k)}$$

I pesi vengono poi normalizzati in modo che i vettori $\mathbf{x} = [x_1, \dots, x_n]$ e $\mathbf{y} = [y_1, \dots, y_n]$ abbiano norma unitaria.

Algoritmo HITS

In notazione matriciale si ha:

$$\mathbf{x}^{(k)} = A^T \mathbf{y}^{(k-1)} \qquad \text{e} \qquad \mathbf{y}^{(k)} = A\mathbf{x}^{(k)}$$

cioè

$$\mathbf{x}^{(k)} = c_k A^T A \mathbf{x}^{(k-1)} \qquad \text{e} \qquad \mathbf{y}^{(k)} = c'_k A A^T \mathbf{y}^{(k-1)}$$

Quindi, l'algoritmo HITS è il metodo delle potenze applicato alle matrici A^TA e AA^T e i vettori contenenti gli indici di authority e gli indici di hub sono gli autovettori dominanti di queste matrici, rispettivamente.

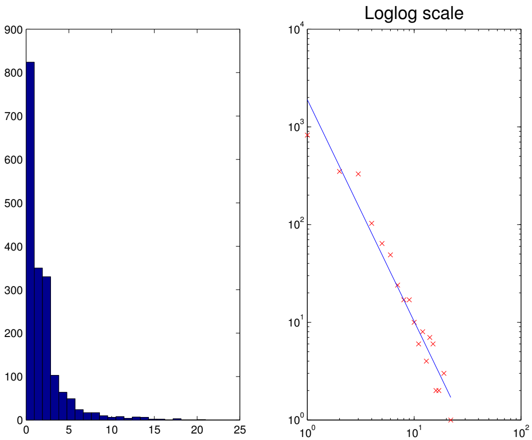
Power low degree distribution

Una rete è scale-free se la distribuzione dei gradi dei suoi nodi segue una legge di potenza, cioè la frazione $F(k)$ dei nodi che hanno grado k si comporta come

$$F(k) \sim ck^p$$

dove c è una costante di normalizzazione e p è un parametro il cui valore, in genere, varia tra -2 e -3.

Indici di Grafo



d_{min} = 0 d_{max} = 21

p = −2.27 – R = 0.98

Indici di Grafo

Diametro e Distanza media

Il diametro di una rete è definito come la massima distanza tra due nodi nella rete, cioè

$$\mathrm{diam}(G) = \max_{i,j \in V} \mathrm{dist}(i,j)$$

La distanza media è la media delle distanze al variare delle coppie di nodi connessi:

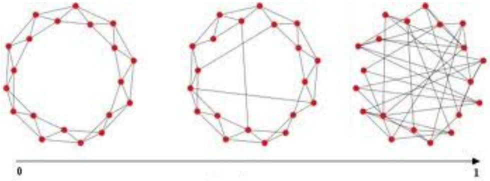
$$\ell(G) = \frac{1}{N} \sum_{(i,j) \in \mathcal{C}} \mathrm{dist}(i,j)$$

dove $\mathcal{C} = \{(i,j) : \mathrm{dist}(i,j) < \infty\}$.

Effetto small-world

Si dice che una rete presenta un effetto small-world se il valore di $\ell(G)$ cresce al più come $\log n$.

Modello di una rete small-world:



Indici di Grafo

Indice di Clustering

È la media degli indici di clustering di tutti i nodi

$$\text{clust}(G) = \frac{1}{n}\sum_{i=1}^{n} C(i)$$

dove $C(i)$ è la probabilità che due nodi connessi a uno stesso nodo siano a loro volta connessi tra loro.

$$C(i) = \frac{t_i}{T_i} = \frac{(A + A^T)^3_{ii}}{2[\text{deg\_tot}(i)\,(\text{deg\_tot}(i) - 2[A^2]_{ii})]}$$

Indice di Estrada

È definito come la somma delle subgraph centralities di ogni nodo:

$$EE(G) = \sum_{i=1}^{n} \left[e^{A}\right]_{ii} = \text{Traccia}\left(e^{A}\right) = \sum_{i=1}^{n} e^{\lambda_i}$$

Reti del Software

I moderni sistemi software sono composti da svariate unità elementari, chiamate moduli, legate tra loro in modo da svolgere specifici compiti.

In particolare, nei sistemi object-oriented (OO) i moduli sono le classi e il legame tra loro è dato da relazioni quali ereditarietà o dipendenza.

Recentemente lo studio di tali sistemi software ha potuto beneficiare della teoria delle reti complesse.

Netbeans

Il sistema software preso in considerazione è Netbeans.

Questo sistema è scritto in Java e le sue classi sono contenute in file sorgenti chiamati Compilation Units (CU).
La maggior parte delle CU contiene soltanto una classe e solo una piccola percentuale ne contiene più di una.

Lo scopo del lavoro in corso è quello di verificare il diffondersi di bug all'interno del sistema e dare quindi una misura della fallibilità del software.

Software graphs

Il vantaggio di utilizzare un sistema software quale Netbeans sta nel fatto che sia il codice sorgente di numerose versioni sia i dati riguardanti i bug sono completamente disponibili.

In particolare, le informazioni sui bug sono date a livello di CU ed è quindi naturale associare al sistema software una rete complessa nel modo seguente:

- i nodi sono le CU
- la CU i e la CU j sono connesse con un arco orientato se nella seconda è presente una classe che dipende o che eredita i moduli dalla prima.

Software graphs

È in corso un lavoro di analisi su un insieme di reti corrispondenti a diversi parti del software con un numero di nodi variabile: da un minimo di 2 nodi (un arco) a un massimo di 44581 nodi (189646 archi).

È allora facile vedere come la scelta del giusto algoritmo sia cruciale nell'analisi di queste reti.

Problemi

- La rete è orientata (la matrice non è simmetrica);
- Sono presenti dei nodi disconnessi ($\sim 10\%$);
- Sono presenti dei cicli.

Algoritmi

Indici di Hub e di Authority

- HITS [Kleinberg, 1999]
- HITS esponenziale [Farahat et al., 2006]

  Consiste nell'applicare l'algoritmo HITS sostituendo la matrice A con la matrice $\left(e^A - I\right)$
- Hub centrality e Authority centrality [Benzi et al., submitted]

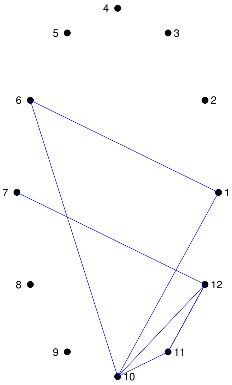
  Consiste nel calcolare le subgraph centralities della matrice

$$\begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}$$

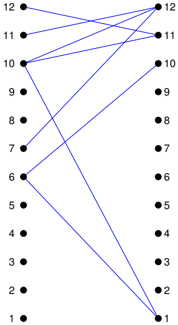
cioè la matrice di adiacenza associata al grafo bipartito.

Risultati numerici

Nodi: 12 – Archi:18



Grafo associato alla matrice



Grafo bipartito associato alla matrice

Risultati numerici

Indici di Hub

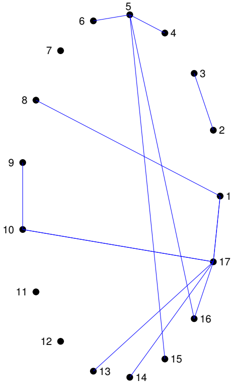
Metodo	HITS	Exp_HITS	Hub_centr
1°	10	10	10
2°	7	6	6
3°	11	7	7
4°	6	11	11
5°	12	12	12

Indici di Authority

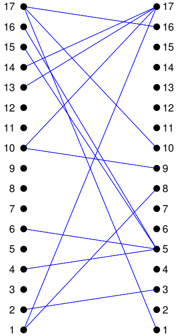
Metodo	HITS	Exp_HITS	Auth_centr
1°	12	12	12
2°	1	11	1
3°	11	1	11
4°	10	10	10
5°	2	2	2

Risultati numerici

Nodi: 17 – Archi:14



Grafo associato alla matrice



Grafo bipartito associato alla matrice

Risultati numerici

Indici di Hub

Metodo	HITS	Exp_HITS	Hub_centr
1°	1	17	17
2°	10	10	1
3°	13	1	10
4°	14	13	14
5°	4	14	13

Indici di Authority

Metodo	HITS	Exp_HITS	Auth_centr
1°	17	17	17
2°	8	10	5
3°	9	1	1
4°	5	16	10
5°	1	9	16

Algoritmi

Distanza media

- Algoritmo "base"
- Burning algorithm
- Dijkstra's algorithm

Burning algorithm

Consta di due passi:

- Riordinare i nodi in ordine topologico, cioè in modo che
$$(i,j) \in E \longrightarrow i < j.$$
- Trovare la distanza minima tramite un *pulling* algorithm, cioè calcolando le distanze minime a ritroso fino ad arrivare al nodo considerato.

Dijkstra's algorithm

Consiste nel dividere i nodi in due gruppi: quelli permanenti e quelli temporanei.

I nodi permanenti sono quelli per cui la distanza è stata calcolata e quelli temporanei sono i nodi per i quali bisogna calcolarla.

Per un dato nodo i si inizializza un vettore d con valori uguali a infinito. Si "percorre" poi la rete a partire dal nodo i seguendo ogni arco esistente. Qundo si arriva al nodo j, si sostuisce il valore di $d(j)$ con il numero di passi effettuati per arrivare a quel nodo e si etichetta il nodo j come permanente.

Alla fine del calcolo nel vettore d saranno memorizzate le distanze dal nodo i a tutti gli altri nodi.

Per ottenere tutte le distanze si devono considerare tutti i nodi uno per volta.

Risultati numerici

Distanza media

Nodi	Archi	$\ell(G)\_1$	$\ell(G)\_2$	Time_1	Time_2
166	284	$2,48$	$2,48$	$0,25$	$0,08$
278	519	$2,55$	$2,55$	$1,37$	$0,15$
444	1146	$3,16$	$3,16$	$6,84$	$0,86$
576	1342	$3,63$	$3,63$	$17,6$	$1,26$
1275	3387	$4,35$	$4,35$	272	16
1447	2831	$3,69$	$3,69$	479	6

Approssimazione tramite formule di quadratura

Supponiamo di volere trovare delle approssimazioni per la forma quadratica

$$v^T f(A) v$$

Se A è simmetrica consideriamo la sua fattorizzazione spettrale $A = Q \Lambda Q^T$, dove Q è una matrice $n \times n$ contenente gli autovettori di A e Λ è una matrice diagonale contenente gli autovalori di A. Allora si ha:

$$v^T f(A) v = v^T Q f(\Lambda) Q^T v = \sum_{i=1}^{n} f(\lambda_i) {w_i}^2$$

dove $w = Q^T v$. Questa somma può essere interpretata come un integrale di Stieltjes

$$v^T f(A) v = \int_{\lambda_1}^{\lambda_n} f(\lambda) d\mu(\lambda) \qquad (1)$$

Approssimazione tramite formule di quadratura

Questo integrale può essere approssimato mediante una formula di quadratura del tipo

$$\int_{\lambda_1}^{\lambda_n} f(\lambda) d\mu(\lambda) = \sum_{j=1}^{N} w_j f(s_j) + \sum_{k=1}^{M} v_k f(z_k) + R[f].$$

Si ha:

- $M = 0$ per la regola di Gauss;
- $M = 1$ ($z_1 = \lambda_1$ oppure $z_1 = \lambda_n$) per la regola di Gauss-Radau;
- $M = 2$ ($z_1 = \lambda_1$ e $z_2 = \lambda_n$) per la regola di Gauss-Lobatto.

Sotto opportune ipotesi sulla funzione f, le formule di quadratura forniscono delle approssimazioni per difetto e/o per eccesso per l'integrale (1).

Metodo di bidiagonalizzazione di Lanczos

Per esempio, supponiamo di voler calcolare l'elemento di posto (i, j) di $f(A)$ tramite la formula di Gauss allora la forma quadratica prende la forma [Golub and Meurant, 2010]

$$e_i^T f(A) e_j = \sum_{j=1}^{N} w_j f(s_j) = e_1^T f(J_N) e_1$$

dove

$$J_N = \begin{bmatrix} \alpha_1 & \beta_1 & & & \\ \beta_1 & \alpha_2 & \beta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \beta_{N-2} & \alpha_{N-1} & \beta_{N-1} \\ & & & \beta_{N-1} & \alpha_N \end{bmatrix}$$

GRAZIE