Calcolo di indici di centralità di reti sociali

E. Bozzo, D. Fasino, M. Franceschet
Università di Udine

Due Giorni di Algebra Lineare Numerica
Genova, 15–16 Febbraio 2012

Riferimenti

M.E.J. Newman.
Networks. An introduction.
Oxford University Press, NY, 2010.

K. Stephenson and M. Zelen.
Rethinking centrality: Methods and examples,
Social Networks, 11 (1989), 1–37.

M.E.J. Newman
A measure of betweenness centrality based on random walks,
Social Networks, 27 (2005), 39–54.

A. Ghosh, S. Boyd and A. Saberi
Minimizing Effective Resistance of a Graph
SIAM Review, 50 (2008), 37–66.

Reti Sociali

In una rete sociale i nodi rappresentano generalmente persone, chiamate in gergo attori, mentre gli archi interazioni sociali, chiamate legami.
Attualmente Facebook è la rete sociale per antonomasia, ma in sociologia vi è una tradizione consolidata nello studio empirico delle reti.

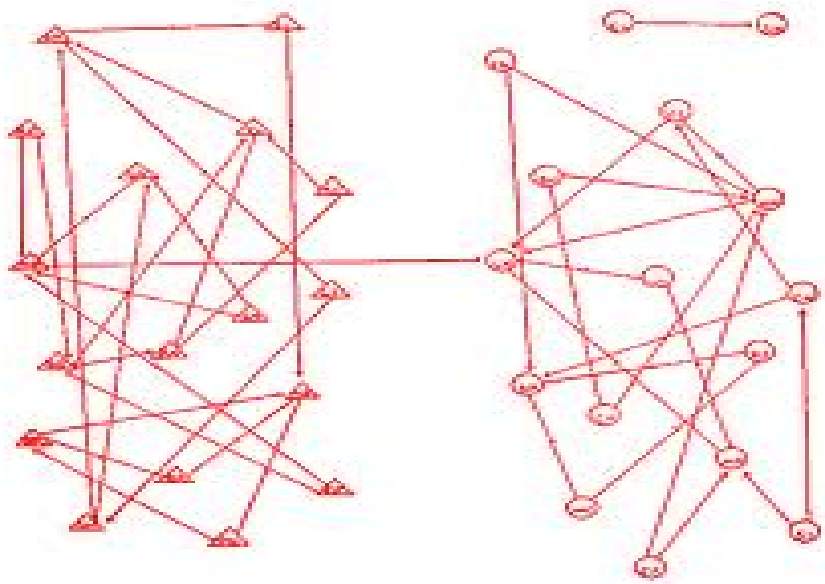


Figure: Amicizie tra maschi e femmine di una classe elementare (1933)

Reti Sociali

Altri esempi notevoli dagli articoli citati

- Attori: babbuini Gelada. Legami: interazioni non agonistiche (generalmente lo spulciamento) tra gli adulti.

Reti Sociali

Altri esempi notevoli dagli articoli citati

- Attori: babbuini Gelada. Legami: interazioni non agonistiche (generalmente lo spulciamento) tra gli adulti.
- Attori: maggiori famiglie fiorentine del ’500. Legami: matrimoni tra componenti delle famiglie.

Reti Sociali

Altri esempi notevoli dagli articoli citati

- Attori: babbuini Gelada. Legami: interazioni non agonistiche (generalmente lo spulciamento) tra gli adulti.
- Attori: maggiori famiglie fiorentine del ’500. Legami: matrimoni tra componenti delle famiglie.
- Attori: omosessuali con diagnosi di AIDS della regione di Los Angeles. Legami: rapporti sessuali.

Reti Sociali

Altri esempi notevoli dagli articoli citati

- Attori: babbuini Gelada. Legami: interazioni non agonistiche (generalmente lo spulciamento) tra gli adulti.
- Attori: maggiori famiglie fiorentine del ’500. Legami: matrimoni tra componenti delle famiglie.
- Attori: omosessuali con diagnosi di AIDS della regione di Los Angeles. Legami: rapporti sessuali.
- Attori: studiosi di reti. Legami: indicano che i due studiosi sono coautori di un lavoro scientifico.

Reti Sociali

Altri esempi notevoli dagli articoli citati

- Attori: babbuini Gelada. Legami: interazioni non agonistiche (generalmente lo spulciamento) tra gli adulti.
- Attori: maggiori famiglie fiorentine del ’500. Legami: matrimoni tra componenti delle famiglie.
- Attori: omosessuali con diagnosi di AIDS della regione di Los Angeles. Legami: rapporti sessuali.
- Attori: studiosi di reti. Legami: indicano che i due studiosi sono coautori di un lavoro scientifico.

Si tratta di reti indirette. Ne seguito trattiamo per semplicità il caso non pesato. L’estensione al caso pesato non presenta difficoltà.
L’estensione alle reti dirette si presenta invece problematica.

Indici di centralità

Gli indici di centralità più usati per reti sociali sono tre.

- Grado. Misura la popolarità di un attore. Il fatto che sia immediatamente disponibile lo rende un termine di paragone fondamentale per ogni altro indice. Un indice alternativo può risultare utile perchè conduce anche solo per certi nodi ad una classifica molto differente. Anche se qui non ne trattiamo pensiamo all'approccio alla Google per reti di citazioni. Un lavoro emblematico è
  P. Chen, H. Xie, S. Maslov, and S. Redner.
  Finding scientific gems with Google's PageRank algorithm.
  Journal of Informetrics, 2007.

Indici di centralità

- Closeness. Media delle “distanze” di un nodo dagli altri. Fornisce una misura della rapidità di propagazione dell’informazione da un certo nodo (evidentemente l’informazione potrebbe essere il virus HIV).
- Betweewnness. Media del numero di volte in cui un nodo si trova in un “cammino” tra altri due nodi. È una misura di quanto un attore ha il controllo dell’informazione che scorre tra gli altri (in una rete di coautori un’alta Betweewnness può indicare che l’attore funge da collegamento tra comunità che lavorano su argomenti diversi).

Oltre l'approccio geodesico

Probabilmente nel frame precedente avete inteso "cammino" come cammino di lunghezza minima e "distanza" come la minima tra la lunghezza dei cammini tra i due nodi. Questo approccio, intrapreso inizialmente, è stato successivamente criticato in quanto in una rete sociale è riduttivo pensare che l'informazione scorra solo lungo cammini minimi.
Una delle proposte alternative può essere descritta tramite un'intuitiva analogia elettrica: la rete viene assimilata ad una rete di resistori aventi (nel caso non pesato) resistenza di 1 Ω.

Il Laplaciano

Supponiamo di scegliere due nodi di indici p e q distinti e di chiedere che dal primo entri nella rete una corrente di 1 A e che esca dal secondo. Chiediamoci quali devono essere i potenziali ai nodi per mantenere stabile tale flusso di corrente. Ovviamente per ogni nodo diverso da p e da q la somma delle correnti entranti e uscenti dal nodo deve essere nulla. Quindi, per ogni nodo i della rete

$$\sum_{j \backsim i}(v_i - v_j) = d_i v_i - \sum_{j \backsim i} v_j = \begin{cases} 0, & p \neq i \neq q; \\ 1, & i = p; \\ -1, & i = q. \end{cases}$$

Se D è la matrice diagonale dei gradi dei nodi e A è la matrice di adiacenza della rete otteniamo

$$(D - A)v = Lv = e_p - e_q.$$

La matrice $L = D - A$ è detta Laplaciano della rete.

Proprietà del Laplaciano

Il Laplaciano di una rete indiretta $\mathcal{G}$ con n nodi e m archi

- è una matrice simmetrica di ordine n;
- è una matrice singolare in quanto $Le = 0$;
- è una matrice semidefinita positiva in quanto $L = BB^T$ essendo B una matrice $n \times m$ tale che se l'arco l connette i nodi i e j allora $B(i, l) = 1$, $B(j, l) = -1$ e $B(i, k) = 0$ altrimenti;
- gli autovalori di L possono quindi essere ordinati come segue

$$0 \leq \lambda_2 \leq \ldots \leq \lambda_n$$

e λ_2 è chiamato valore di Fiedler o connettività algebrica della rete;

Proprietà del Laplaciano (continua)

- vale il Matrix Tree Theorem che può essere espresso nella forma

$$t(\mathcal{G}) = \frac{1}{n} \prod_{i=2}^{n} \lambda_i$$

essendo $t(\mathcal{G})$ il numero di alberi di supporto della rete;

- dal punto precedente otteniamo che se il valore di Fiedler è nullo allora la rete non è connessa ma più in generale si dimostra che la molteplicità di λ_2 coincide con il numero di componenti connesse;
- per una rete connessa l'autovettore normalizzato v_2 associato a λ_2 è chiamato vettore di Fiedler della rete e

$$\lambda_2 = \min_{\|v\|=1, v \perp e} v^T L v = v_2^T L v_2 = v_2^T B B^T v_2 = \sum_{j \backsim i} (v_2(i) - v_2(j))^2$$

e questo suggerisce che nodi adiacenti debbano avere componenti del vettore di Fiedler vicine.

Resistance distance

D'ora in avanti assumiamo di lavorare con reti connesse. Evidentemente il sistema $Lv = e_p - e_q$ ha soluzione in quanto il vettore dei termini noti è ortogonale ad e, inoltre due generiche soluzioni differiscono per un multiplo di e (fisicamente contano solo le differenze di potenziale). Possiamo allora scegliere

$$w = L^+(e_p - e_q).$$

Definiamo

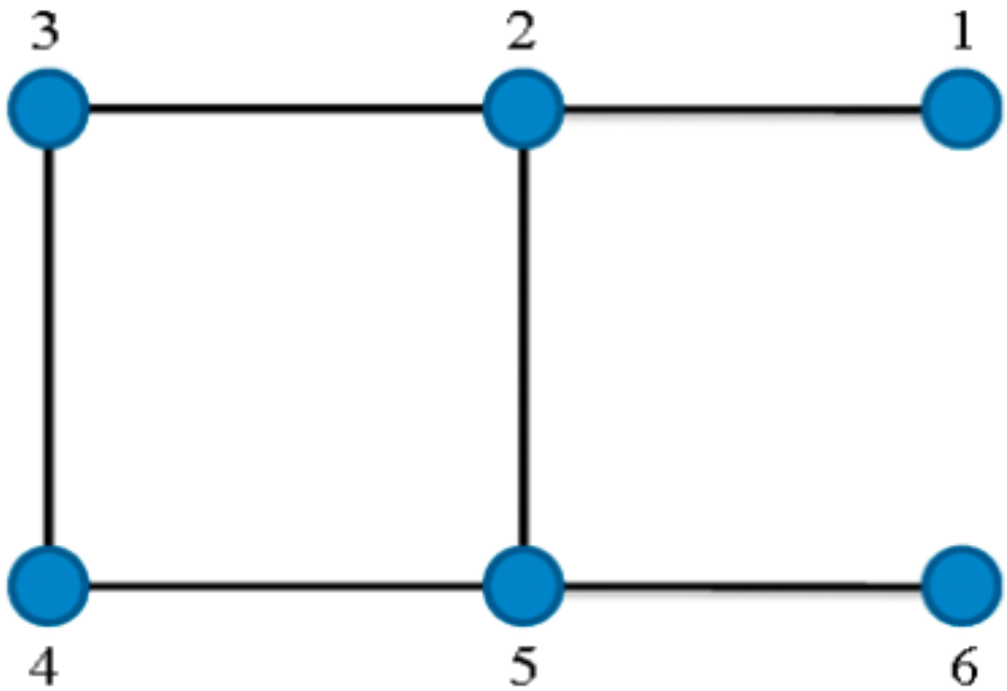
$$R_{pq} = w_p - w_q = (e_p - e_q)^T L^+(e_p - e_q) = L^+(p,p) + L^+(q,q) - 2L^+(p,q)$$

resistance distance dei nodi p e q. Si dimostra che

- se $p \neq q$ allora $R_{pq} > 0$ (L^+ è definita positiva sullo spazio ortogonale ad e);
- risulta $R_{pq} \leq R_{ps} + R_{sq}$ il che giustifica l'utilizzo del termine distanza.

Resistance distance

R_{pq} risulta minore o uguale alla distanza geodesica tra p e q. Per esempio nel grafo seguente la distanza geodesica tra il nodo 2 ed il nodo 5 è 1 mentre la resitance distance è $0.75 = \frac{1}{1+\frac{1}{3}}$.



Calcolo della closeness centrality

A questo punto siamo in grado di ottenere una semplice formula per la closeness centrality. Infatti

$$C(p) = \frac{1}{n}\sum_{q=1}^{n} R_{pq} = \frac{1}{n}\sum_{q=1}^{n}\left(L^+(p,p) + L^+(q,q) - 2L^+(p,q)\right),$$

e siccome $Le = 0$ implica $L^+e = 0$ (una matrice simmetrica e la sua pseudoinversa di Moore e Penrose hanno lo stesso nucleo)

$$C(p) = L^+(p,p) + \frac{1}{n}Tr(L^+)$$

e dunque ci riconduciamo al problema del calcolo degli elementi diagonali di L^+. Tenuto conto che $L^+ = (L + \frac{1}{n}ee^T)^{-1} - \frac{1}{n}ee^T$ ci siamo ripromessi di ottenere delle stime con un approccio alla Golub-Meurant per il calcolo di forme bilineari del tipo $u^T f(A)u$ che è stato applicato con con buoni risultati alla stima degli elementi diagonali dell'esponenziale della matrice di adiacenza (work in progress...).

Calcolo della betweenness centrality

La proposta di Newman è di calcolare la quantità di corrente che attraversa un nodo scelta una coppia di nodi di “ingresso” e “uscita” p e q come segue

$$B^{(pq)}(i) = \frac{1}{2}\sum_{j=1}^{n} A(i,j)|(e_i - e_j)^T L^+ (e_p - e_q)|$$

e poi calcolare la media dei valori ottenuti al variare di p e q

$$B(i) = \frac{\sum_{p<q} B^{(pq)}(i)}{(1/2)n(n-1)}.$$

Calcolo della betweenness centrality

I costi di tutto questo sono notevoli:

- L^+ deve essere calcolata completamente ed è generalmente piena;
- il calcolo dei $B(i)^{(pq)}$ costa $O(mn^2)$;
- il calcolo dei $B(i)$ costa $O(n^3)$.

Per reti di grandi dimensioni (Facebook...) sono stati proposti algoritmi randomizzati che eseguono i calcoli solo per un certo numero di coppie p e q scelte a caso. Si veda per esempio
U. Brandes e D. Fleischer, Centrality measures based on current flow, STACS 2005.

Calcolo della betweenness centrality

La nostra proposta è quella di utilizzare un certo numero di autocoppie del Laplaciano per ottenere delle approssimazioni. Risulta

$$\widetilde{B}_r^{(pq)}(i) = \frac{1}{2}\sum_{j=1}^{n} A(i,j)\left|\sum_{k=2^r} \frac{1}{\lambda_k}(e_i - e_j)^T v_k v_k^T (e_p - e_q)\right|.$$

Questo contiene le richieste di memoria anche se il calcolo dei $\widetilde{B}_r^{(pq)}(i)$ costa $O(mn^2 r)$.

Calcolo della betweenness centrality

Nel caso $r = 2$ comunque

$$\begin{aligned}\widetilde{B}_2^{(pq)}(i) &= \frac{1}{2}\sum_{j=1}^{n} A(i,j)\left|\frac{1}{\lambda_2}(e_i - e_j)^T v_2 v_2^T (e_p - e_q)\right| \\ &= \frac{|v_2(p) - v_2(q)|}{2\lambda_2}\sum_{j=1}^{n} A(i,j)|v_2(i) - v_2(j)|\end{aligned}$$

il che porta i costi di calcolo ad $O(m)$, oltre naturalmente al calcolo di v_2.

Domande

- Per il calcolo del vettore di Fiedler quanto gli algoritmi multilivello sono meglio di `eigs`?
- Come si ripercuote l'errore nell'approssimazione di L^+ su quello sulle centralità?
- ...

A cosa servono gli indici di centralità?

Our view is that centrality is only a descriptive property of a network. An area of future research should be concerned with innovative uses of centralities to describe how networks may change over time or to determine the consequences of new scenarios when nodes or lines are added or deleted.

Karen Stephenson and Marvin Zelen