Motori di ricerca e reti complesse

Bruno **Iannazzo**, Università di **Perugia**

Contare chi conta

Il modello del motore di ricerca

Nel mondo c'è solo una cosa peggiore che si parli di noi,
e cioè che non se ne parli affatto.

The portrait of Dorian Gray, *Oscar Wilde*

Motori di ricerca

Un modo per trovare le pagine in Internet è tramite parole chiave (alternative: seguire i link o usare una web directory: Yahoo!, DMoz)

Una lista di tutte le pagine che contengono una parola chiave è inutile.

Un utente vorrebbe trovare le pagine che gli interessano senza scorrerne molte.

Un modo per ordinare le pagine è tramite la **rilevanza/importanza**.

Motori di ricerca

Problema linguistico

Cos'è l'importanza?

Motori di ricerca

Problema linguistico

Cos'è l'importanza?

Problema: modellistico/pratico

Come far funzionare le cose? Dare una definizione matematica di importanza e un modo di calcolarla (o approssimarla) in un tempo ragionevole.

Relazione con il mondo reale

$$\begin{tikzcd}[column sep=huge] \text{Mondo reale} \arrow[r, bend left, "\text{Modell(izz)azione}"] & \text{Matematica} \arrow[l, bend left, "\text{Applicazione}"] \end{tikzcd}$$

Motori di ricerca

Primo tentativo: **importanza soggettiva** (ad es. altavista)

La rilevanza rispetto a una parola chiave è data da una serie di proprietà

- numero di volte in cui la parola appare nella pagina;
- posizione dove la parola compare (titolo, meta tag, inizio del testo).

Motori di ricerca

Primo tentativo: **importanza soggettiva** (ad es. altavista)

La rilevanza rispetto a una parola chiave è data da una serie di proprietà

- numero di volte in cui la parola appare nella pagina;
- posizione dove la parola compare (titolo, meta tag, inizio del testo).

```
<head>
<title>Perugia</title>
<meta name="keywords" content="Perugia">
</head>
<body>
Perugia
...
</body>
```

Motori di ricerca

Primo tentativo: **importanza soggettiva** (ad es. altavista)

La rilevanza rispetto a una parola chiave è data da una serie di proprietà

- numero di volte in cui la parola appare nella pagina;
- posizione dove la parola compare (titolo, meta tag, inizio del testo).

Problema

Quest'approccio è fallito! Spiegare perché.

Motori di ricerca

Primo tentativo: **importanza soggettiva** (ad es. altavista)

La rilevanza rispetto a una parola chiave è data da una serie di proprietà

- numero di volte in cui la parola appare nella pagina;
- posizione dove la parola compare (titolo, meta tag, inizio del testo).

Quest'approccio è diventato inutilizzabile a causa dello spam (siti web farlocchi con molte parole chiave non legate al contenuto commerciale)

Motori di ricerca

Nuovo approccio: **importanza oggettiva** (ad es Google)

La rilevanza rispetto a una parola chiave è data dall'importanza della pagina che la contiene

Di nuovo: cos'è l'importanza? (E come ridurre lo spam?)

Verso una definizione di importanza

Assegnare un **numero** (etichetta) ad ogni pagina nel web. Siano $\{1, \dots, N\}$ le pagine.

L'importanza sarà una **funzione** $p : \{1, \dots, N\} \to \mathbb{R}$ tale che p_i è un numero che rappresenta l'importanza della pagina i.

Si fanno delle assunzioni di comodo:

- p_i sia reale: per non avere problemi, ad esempio, quando vengono fuori radici quadrate;
- $p_i \geqslant 0$: ragioniamo solo su importanze non negative;
- $\sum_{i=1}^{n} p_i = 1$: l'importanza è relativa, allora facciamo in modo che la somma di tutte le importanze sia 1.

Si osservi che N è **graaande**. Il 24 ottobre 2016, Google aveva indicizzato circa 50 miliardi di pagine.
(fonte `http://www.worldwidewebsize.com/`)

Verso una definizione di importanza

Idea: **usare i link!**

Una pagina è importante se riceve molti link

Verso una definizione di importanza

Idea: **usare i link!**

Una pagina è importante se riceve molti link

Vicolo cieco: uno spammer può creare un gran numero di pagine che linkano la sua pagina

Verso una definizione di importanza

Un'idea per superare questo fatto è di porre

Una pagina è importante se riceve molti link da pagine importanti

Vicolo cieco: non è la stessa cosa

- ricevere un link da una pagina molto importante che linka milioni di pagine
- ricevere un link da una pagina un po' importante che linka poche pagine

Una definizione di importanza

L'importanza di una pagina è data dall'importanza delle pagine che hanno un link verso di essa, pesata per il numero di pagine che esse linkano

Una definizione di importanza

L'importanza di una pagina è data dall'importanza delle pagine che hanno un link verso di essa, pesata per il numero di pagine che esse linkano

Si traduce facilmente in un'equazione! Siamo vicini al modello matematico!

Una definizione di importanza

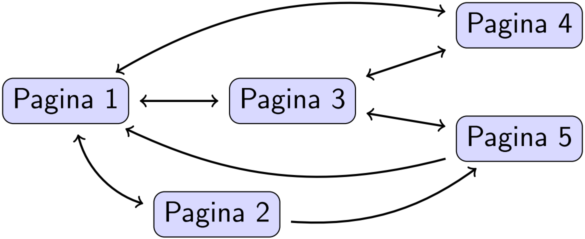
L'importanza di una pagina è data dall'importanza delle pagine che hanno un link verso di essa, pesata per il numero di pagine che esse linkano

Sia u_j il numero di link uscenti da j e $\mathcal{E}_k$ l'insieme delle pagine che linkano la pagina k (link entranti)

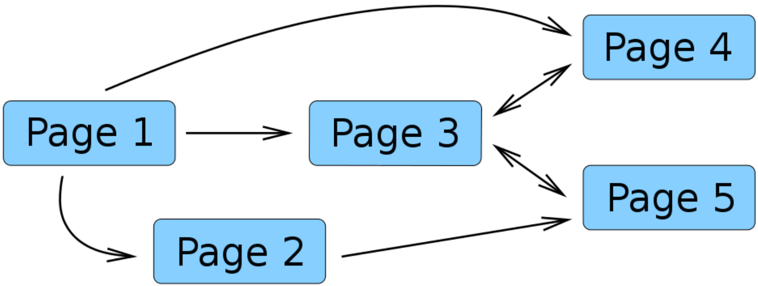
$$p_k = \sum_{j \in \mathcal{E}_k} \frac{p_j}{u_j}, \quad p_k \geqslant 0, \quad \sum_{k=1}^{N} p_k = 1$$

La definizione sembra ricorsiva $\rightarrow$ l'equazione è implicita: non permette di calcolare l'importanza direttamente.

Una rete giocattolo



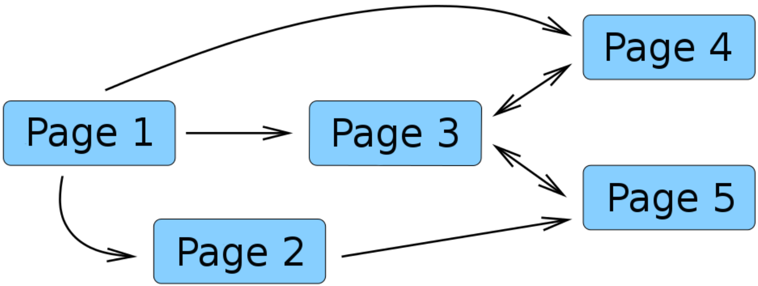
Una rete giocattolo



Assumiamo che ogni pagina linki "Page 1" e se stessa
$(u_1 = 4, u_2 = 3, u_3 = 4, u_4 = 3, u_5 = 3)$
L'equazione dell'importanza è per il nodo 1

$$p_1 = \frac{p_1}{u_1} + \frac{p_2}{u_2} + \frac{p_3}{u_3} + \frac{p_4}{u_4} + \frac{p_5}{u_5}$$

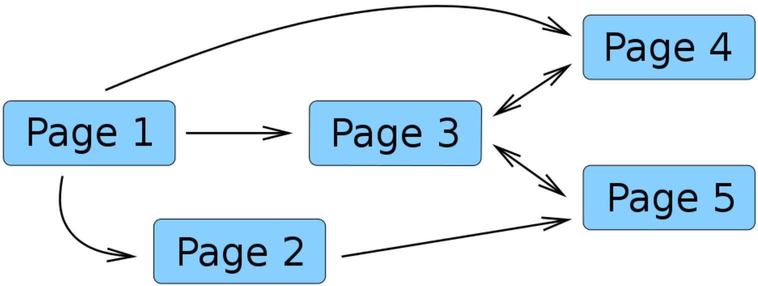
Una rete giocattolo



Assumiamo che ogni pagina linki “Page 1” e se stessa
($u_1 = 4, u_2 = 3, u_3 = 4, u_4 = 3, u_5 = 3$)
L’equazione dell’importanza è per il nodo 2

$$p_2 = \frac{p_1}{u_1} + \frac{p_2}{u_2}$$

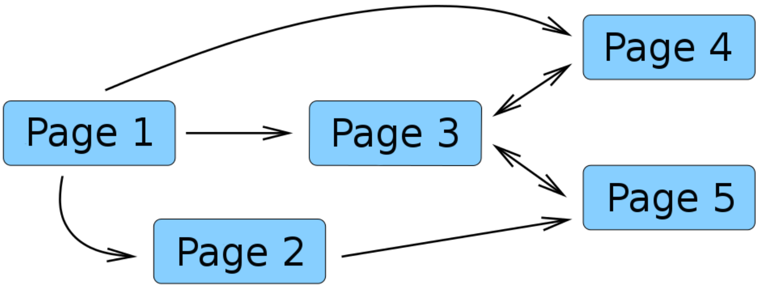
Una rete giocattolo



Assumiamo che ogni pagina linki "Page 1" e se stessa
($u_1 = 4, u_2 = 3, u_3 = 4, u_4 = 3, u_5 = 3$)
L'equazione dell'importanza è per il nodo 3

$$p_3 = \frac{p_1}{u_1} + \frac{p_3}{u_3} + \frac{p_4}{u_4} + \frac{p_5}{u_5}$$

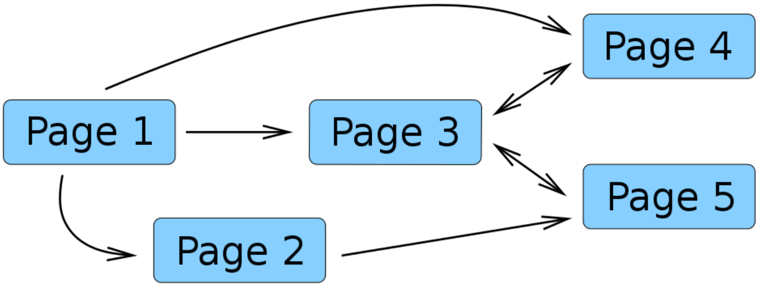
Una rete giocattolo



Assumiamo che ogni pagina linki "Page 1" e se stessa
($u_1 = 4, u_2 = 3, u_3 = 4, u_4 = 3, u_5 = 3$)
L'equazione dell'importanza è per il nodo 4

$$p_4 = \frac{p_1}{u_1} + \frac{p_3}{u_3} + \frac{p_4}{u_4}$$

Una rete giocattolo



Assumiamo che ogni pagina linki "Page 1" e se stessa
($u_1 = 4, u_2 = 3, u_3 = 4, u_4 = 3, u_5 = 3$)
L'equazione dell'importanza è per il nodo 5

$$p_5 = \frac{p_2}{u_2} + \frac{p_3}{u_3} + \frac{p_5}{u_5}$$

Problema matematico (e anche un po' informatico)

Studiare l'equazione

$$p_k = \sum_{j \in \mathcal{E}_k} \frac{p_j}{u_j}, \quad p_k \geqslant 0, \quad \sum_{k=1}^{N} p_k = 1$$

- L'equazione ha **sempre** soluzione?
- **Quante soluzioni** esistono?
- È possibile **calcolare praticamente** questa soluzione (poiché N è enorme)?

Ognuna di queste domande ha una risposta!

Un'altra via per l'importanza

Un modello standard per il Web è un **grande grafo diretto**

- nodi $\longleftrightarrow$ pagine
- archi orientati $\longleftrightarrow$ link

È quello che abbiamo visto prima.

Un'altra via per l'importanza

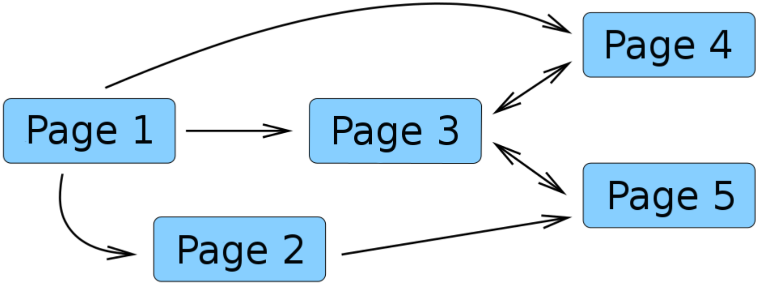
Un modello standard per il Web è un **grande grafo diretto**

- nodi $\longleftrightarrow$ pagine
- archi orientati $\longleftrightarrow$ link

È quello che abbiamo visto prima.

Dal grafo con N nodi si riesce a costruire una tabella (**matrice di adiacenza**) con N righe e N colonne

Esempio



Assumiamo anche che ogni pagina linki “Page 1” e se stessa

Esempio

Inseriamo un 1 nell'elemento che si trova nella riga i e nella colonna j se c'è un link tra la pagina i e la pagina j.

La matrice di adiacenza per l'esempio precedente è

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{bmatrix}$$

Un'altra via per l'importanza

Ora modifichiamo la matrice nel seguente modo:

dividiamo ogni elemento di una riga della matrice di adiacenza A per la somma della riga ottenendo la matrice **stocastica** $G = (g_{ij})$ (la matrice di Google), tale che

$$g_{ij} = \frac{a_{ij}}{\sum_{j=1}^{N} a_{ij}}$$

Osserviamo che $\sum_{j=1}^{N} a_{ij} = \nu_i$ (numero di link uscenti)

Una matrice $M = (m_{ij})$ è detta stocastica se

- M è nonnegativa, i.e. $m_{ij} \geqslant 0$ per ogni i, j
- M ha somma per riga 1

In particolare, $Me = e$, dove $e = [1 \;\; 1 \;\; \cdots \;\; 1]^T$ è un autovettore.

Un'altra via per l'importanza

Matrice di adiacenza $\rightarrow$ matrice di Google

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{bmatrix} \rightarrow G = \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 \\ 1/4 & 0 & 1/4 & 1/4 & 1/4 \\ 1/3 & 0 & 1/3 & 1/3 & 0 \\ 1/3 & 0 & 1/3 & 0 & 1/3 \end{bmatrix}$$

Un'altra via per l'importanza

Mi accorgo che gli elementi della prima riga di G sono del tipo $1/u_1$ se c'è un link oppure 0 se non c'è un link e lo stesso vale per le altre righe.

$$G = \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 \\ 1/4 & 0 & 1/4 & 1/4 & 1/4 \\ 1/3 & 0 & 1/3 & 1/3 & 0 \\ 1/3 & 0 & 1/3 & 0 & 1/3 \end{bmatrix} = \begin{bmatrix} 1/u_1 & 1/u_1 & 1/u_1 & 1/u_1 & 0 \\ 1/u_2 & 1/u_2 & 0 & 0 & 1/u_2 \\ 1/u_3 & 0 & 1/u_3 & 1/u_3 & 1/u_3 \\ 1/u_4 & 0 & 1/u_4 & 1/u_4 & 0 \\ 1/u_5 & 0 & 1/u_5 & 0 & 1/u_5 \end{bmatrix}$$

Un'altra via per l'importanza

Ora moltiplico la matrice G^T per un vettore p generico (di n elementi) e uguaglio il risultato a p

$$\begin{bmatrix} 1/4 & 1/3 & 1/4 & 1/3 & 1/3 \\ 1/4 & 1/3 & 0 & 0 & 0 \\ 1/4 & 0 & 1/4 & 1/3 & 1/3 \\ 1/4 & 0 & 1/4 & 1/3 & 0 \\ 0 & 1/3 & 1/4 & 0 & 1/3 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{bmatrix} = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{bmatrix}$$

Un'altra via per l'importanza

Ora moltiplico la matrice G^T per un vettore p generico (di n elementi) e uguaglio il risultato a p

$$\begin{bmatrix} 1/u_1 & 1/u_2 & 1/u_3 & 1/u_4 & 1/u_5 \\ 1/u_1 & 1/u_2 & 0 & 0 & 0 \\ 1/u_1 & 0 & 1/u_3 & 1/u_4 & 1/u_5 \\ 1/u_1 & 0 & 1/u_3 & 1/u_4 & 0 \\ 0 & 1/u_2 & 1/u_3 & 0 & 1/u_5 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{bmatrix} = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{bmatrix}$$

Un'altra via per l'importanza

Ora moltiplico la matrice G^T per un vettore p generico (di n elementi) e uguaglio il risultato a p

$$\begin{bmatrix} \frac{p_1}{u_1} + \frac{p_2}{u_2} + \frac{p_3}{u_3} + \frac{p_4}{u_4} + \frac{p_5}{u_5} \\ \frac{p_1}{u_1} + \frac{p_2}{u_2} \\ \frac{p_1}{u_1} + \frac{p_3}{u_3} + \frac{p_4}{u_4} + \frac{p_5}{u_5} \\ \frac{p_1}{u_1} + \frac{p_3}{u_3} + \frac{p_4}{u_4} \\ \frac{p_2}{u_2} + \frac{p_3}{u_3} + \frac{p_5}{u_5} \end{bmatrix} = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{bmatrix}$$

Ho ottenuto proprio l'equazione dell'importanza!

Pagerank

Tramite le matrici e il prodotto matrice-vettore sono riuscito a scrivere l'equazione dell'importanza nella forma compatta

$$G^T p = p, \quad p_k \geqslant 0, \quad \sum_{k=1}^{N} p_k = 1$$

Pagerank

Tramite le matrici e il prodotto matrice-vettore sono riuscito a scrivere l'equazione dell'importanza nella forma compatta

$$G^T p = p, \quad p_k \geqslant 0, \quad \sum_{k=1}^{N} p_k = 1$$

Una soluzione non nulla dell'equazione $G^T p = p$ è detta **autovettore** della matrice G^T relativo all'autovalore 1.

Chiamiamo **pagerank** p della rete l'autovettore di G^T relativo all'autovalore 1

Problemi

La matematica ci aiuta a risolvere i problemi che possono nascere

- esiste una soluzione?

Problemi

La matematica ci aiuta a risolvere i problemi che possono nascere

- esiste una soluzione? **Sì!**

Problemi

La matematica ci aiuta a risolvere i problemi che possono nascere

- esiste una soluzione? **Sì!**
- la soluzione è unica una volta normalizzata?

Problemi

La matematica ci aiuta a risolvere i problemi che possono nascere

- esiste una soluzione? **Sì!**
- la soluzione è unica una volta normalizzata? **No! Ma con un trucco si riesce a renderla unica!**

Problemi

La matematica ci aiuta a risolvere i problemi che possono nascere

- esiste una soluzione? **Sì!**
- la soluzione è unica una volta normalizzata? **No! Ma con un trucco si riesce a renderla unica!**
- si sa calcolare p?

Problemi

La matematica ci aiuta a risolvere i problemi che possono nascere

- esiste una soluzione? **Sì!**
- la soluzione è unica una volta normalizzata? **No! Ma con un trucco si riesce a renderla unica!**
- si sa calcolare p? **Sì! Usando l'Analisi Numerica (pubblicità).**

Problemi

La matematica ci aiuta a risolvere i problemi che possono nascere

- esiste una soluzione? **Sì!**
- la soluzione è unica una volta normalizzata? **No! Ma con un trucco si riesce a renderla unica!**
- si sa calcolare p? **Sì! Usando l'Analisi Numerica (pubblicità).**
- ed è sufficientemente veloce?

Problemi

La matematica ci aiuta a risolvere i problemi che possono nascere

- esiste una soluzione? **Sì!**
- la soluzione è unica una volta normalizzata? **No! Ma con un trucco si riesce a renderla unica!**
- si sa calcolare p? **Sì! Usando l'Analisi Numerica (pubblicità).**
- ed è sufficientemente veloce? **No! Ma con il trucco di prima si riesce.**

I risultati matematici necessari sono del 1907 (quando non si poteva neanche immaginare la rete).

Una definizione

Una matrice A non negativa (cioè reale con elementi non negativi) è detta **primitiva** se esiste k tale che A^k è positiva (ogni suo elemento è positivo)

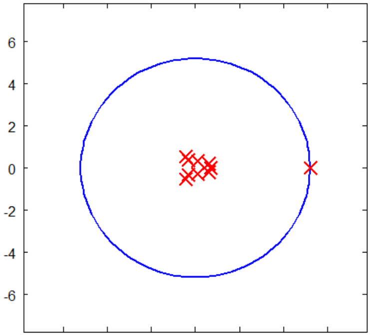
Un teorema del 1907

Teorema (Perron)

Sia $A \in \mathbb{R}^{n\times n}$ primitiva, allora

(a) *esiste un autovalore $\lambda_0 > 0$ di A, a cui corrisponde un autovettore positivo $v > 0$ e non ci sono altri autovettori positivi indipendenti da v;*

(b) *λ_0 è semplice ed è il raggio spettrale di A; inoltre, ogni altro autovalore di A ha modulo strettamente minore di λ_0.*



Un teorema del 1907

Teorema (Perron)

Sia $A \in \mathbb{R}^{n \times n}$ primitiva, allora

(a) *esiste un autovalore $\lambda_0 > 0$ di A, a cui corrisponde un autovettore positivo $v > 0$ e non ci sono altri autovettori positivi indipendenti da v;*

(b) *λ_0 è semplice ed è il raggio spettrale di A; inoltre, ogni altro autovalore di A ha modulo strettamente minore di λ_0.*

λ_0 è detto l'**autovalore di Perron** di A

- 1 è l'autovalore di Perron di G?
- La matrice G è primitiva?

Risposte

- 1 è l'autovalore di Perron di G?

Sì!

Teorema

Sia A stocastica, allora ha raggio spettrale 1.

Risposte

- 1 è l'autovalore di Perron di G?

Sì!

Teorema

Sia A stocastica, allora ha raggio spettrale 1.

- La matrice G è primitiva?

Non sempre!

Si pensi a una rete fatta da due sottoreti che non si linkano a vicenda.

Soluzione brillante

Il "trucco" che sistema tutto è di spezzare l'importanza in due parti

Una frazione c dell'importanza della pagina k è data dai link entranti pesati; la parte restante è fissata da valori v_k.

Perché tutto abbia senso si pone $\sum_{k=1}^{N} v_k = 1$ e il vettore ottenuto è detto vettore di personalizzazione.

Soluzione brillante

Modificare la matrice G e definire una nuova matrice

$$\widetilde{G} = cG + (1-c)ev^T$$

dove e è il vettore di 1, v è un vettore **positivo** arbitrario tale che $v^T e = 1$ (è detto **vettore di personalizzazione**) e $0 < c < 1$ è una costante positiva (nel modello classico $c = 0.85$).

Osserviamo che

- $\widetilde{G}$ è positiva (e quindi primitiva);
- $\widetilde{G}$ è stocastica, cioè $\widetilde{G}e = e$, infatti

Soluzione brillante

Modificare la matrice G e definire una nuova matrice

$$\widetilde{G} = cG + (1-c)ev^T$$

dove e è il vettore di 1, v è un vettore **positivo** arbitrario tale che $v^T e = 1$ (è detto **vettore di personalizzazione**) e $0 < c < 1$ è una costante positiva (nel modello classico $c = 0.85$).

Osserviamo che

- $\widetilde{G}$ è positiva (e quindi primitiva);
- $\widetilde{G}$ è stocastica, cioè $\widetilde{G}e = e$, infatti

$$cGe + (1-c)ev^T e = ce + (1-c)e = e$$

$\widetilde{G}$ **verifica** le ipotesi del teorema di Perron.

Il pagerank corretto

$$\widetilde{G} = cG + (1-c)ev^T$$

Il **pagerank corretto** è l'autovettore di $\widetilde{G}^T$ relativo a 1.

Ora **il modello è cambiato**! Che significato ha $\widetilde{p}$ tale che $\widetilde{G}^T\widetilde{p} = \widetilde{p}$?

Il pagerank corretto

La nuova equazione è

$$\widetilde{p}_k = c \sum_{i \in \mathcal{E}_k} \frac{\widetilde{p}_i}{u_i} + (1-c)v_k$$

che in termini matriciali si può scrivere come

$$p = cG^T p + (1-c)v^T e$$

dove e è il vettore i cui elementi sono tutti uguali a 1.

Si può scegliere $v_i = 1/N$ (dando ad ogni pagina un'importanza una stessa importanza di default) oppure può essere scelto in modo arbitrario (per esempio usando informazioni ulteriori, sottoalgoritmi, eccetera)

Interpretazione probabilistica del pagerank

Una **catena di Markov** (a stati finiti) è un processo stocastico $X_k : \Omega \to \{1, \ldots, N\}$ per $k = 0, 1, 2, \ldots$ tale che

$$P(\{X_k = j\}|\{X_{k-1} = i, X_{k-2} = i_2, \ldots, X_0 = i_k\}) \\ = P(\{X_k = j\}|\{X_{k-1} = i\})$$

La probabilità che al tempo k il processo abbia valore j condizionata con il passato dipende solo dal valore del processo al tempo $k-1$

L'esempio tipico è la passeggiata aleatoria

Se le probabilità non dipendono da k la catena è detta **omogenea**

Interpretazione probabilistica del pagerank

A una catena di Markov omogenea viene associata una matrice $a_{ij} = P(\{X_k = j\}|\{X_{k-1} = i\})$

$A = (a_{ij})$ è stocastica e viene detta **matrice di transizione**

Una probabilità su $\{1, \ldots, N\}$ è una funzione $p = (p_k)$ tale che

$$p_k \geqslant 0, \qquad \sum_{i=1}^{N} p_i = 1$$

Somiglia alla nostra funzione di importanza

Interpretazione probabilistica del pagerank

Teorema

Sia A be la matrice di transizione di una catena di Markov omogenea. Esiste p tale che $A^T p = p$; il vettore p è detto probabilità invariante di A

Il pagerank è una probabilità invariante della matrice di Google!

Interpretazione: se tante persone navigano in Internet seguendo a caso i link in una pagina con la stessa probabilità, dopo un tempo sufficientemente lungo saranno distribuiti approssimativamente come il pagerank

Calcolare p

Ci serve un **algoritmo** per calcolare un autovettore corrispondente all'autovalore di modulo massimo di G^T (o $\widetilde{G}^T$)

Il classico **metodo delle potenze** fa proprio questo!

Idea del metodo delle potenze: partire da un vettore v_0 e applicargli ripetutamente A, ottenendo $v_{k+1} = Av_k$

Sembra stupito, ma con una piccola modifica **funziona bene**.

Il metodo delle potenze

Metodo delle potenze modificato: **normalizzare** il vettore a ogni passo (si evita l'overflow). Dato t_0, viene calcolata la successione

$$\begin{cases} y_{k+1} = At_k \\ t_{k+1} = y_k/\beta_k, \end{cases} \quad k = 0, 1, 2, \ldots$$

dove β_k è uno degli elementi di massimo modulo di y_k

t_k converge a un autovettore di A relativo all'autovalore d massimo modulo λ_1 e $\beta_k \to \lambda_1$, sotto alcune ipotesi.

Il metodo delle potenze

Metodo delle potenze modificato: **normalizzare** il vettore a ogni passo (si evita l'overflow). Dato t_0, viene calcolata la successione

$$\begin{cases} y_{k+1} = At_k \\ t_{k+1} = y_k/\beta_k, \end{cases} \quad k = 0, 1, 2, \ldots$$

dove β_k è uno degli elementi di massimo modulo di y_k

Teorema

Sia $A \in \mathbb{C}^{n\times n}$ con autovalori $\lambda_1, \ldots, \lambda_n$ ordinati per modulo non crescente e tale che $|\lambda_1| > |\lambda_2| \geqslant \cdots \geqslant |\lambda_n|$, allora il metodo delle potenze converge a un autovettore v corrispondente a λ_1 per quasi ogni $t_0 \in \mathbb{C}^n$

Più precisamente, se t_0 ha coordinata non nulla rispetto a v in una base ottenuta completando v a una base di $\mathbb{C}^n$ (con un sottospazio invariante di dimensione $n-1$)

Ipotesi

L'ipotesi $|\lambda_1| > |\lambda_2| \geqslant \cdots \geqslant |\lambda_n|$ è **naturale** per rendere il problema ben posto.

L'ipotesi sul valore iniziale invece è fastidiosa.

Come scegliere t_0?

Ipotesi

L'ipotesi $|\lambda_1| > |\lambda_2| \geqslant \cdots \geqslant |\lambda_n|$ è **naturale** per rendere il problema ben posto.

L'ipotesi sul valore iniziale invece è fastidiosa.

Come scegliere t_0?

Idea naif: a caso, non possiamo essere così sfortunati da beccare un punto in un insieme di misura nulla

Ipotesi

L'ipotesi $|\lambda_1| > |\lambda_2| \geqslant \cdots \geqslant |\lambda_n|$ è **naturale** per rendere il problema ben posto.

L'ipotesi sul valore iniziale invece è fastidiosa.

Come scegliere t_0?

Idea naif: a caso, non possiamo essere così sfortunati da beccare un punto in un insieme di misura nulla

Idea intelligente: a caso, anche se t_0 avesse componente nulla rispetto a v, in aritmetica finita gli errori di arrotondamento produrrebbero una componente non nulla che verrebbe amplificata nei passi successivi.

Uno dei rari casi in cui l'aritmetica finita funziona meglio di quella esatta!

Convergenza del metodo delle potenze

Sappiamo che $\beta_k - \lambda_1 \to 0$ ma si può dimostrare che

$$|\beta_k - \lambda_1| = O(\gamma^k),$$

dove $\gamma = |\lambda_2|/|\lambda_1|$ (**convergenza lineare/esponenziale**)

Ci aspettiamo **convergenza rapida** se $\gamma \ll 1$ e convergenza lenta se $\gamma \approx 1$.

- $\gamma = 0.5$, $\gamma^k < 2.2 \cdot 10^{-16}$ per $k \geqslant 53$,
- $\gamma = 0.99$, $\gamma^k < 2.2 \cdot 10^{-16}$ per $k \geqslant 3588$,

Torniamo al pagerank

Sappiamo che ogni matrice stocastica primitiva ha un autovalore $\lambda_1 = 1$ maggiore degli altri in modulo $\rightarrow$

$\rightarrow$ il metodo delle potenze **si può applicare**

Può succedere che $|\lambda_2| \approx \lambda_1$, quindi la convergenza è lenta.

Torniamo al pagerank

Per la matrice modificata $\widetilde{G} = cG + (1-c)ev^T$, con $c \approx 0.85$ si può dimostrare che l'autovalore dominante è 1 e che $|\lambda_2| < 0.85$ $\rightarrow$ **convergenza rapida**

Teorema (Brauer 1952)

Sia G una matrice i cui autovalori sono $\lambda_1, \lambda_2, \dots, \lambda_n$ e sia w un autovettore relativo a λ_1, $x \in \mathbb{C}^n$, allora gli autovalori della matrice $G + wx^*$ sono $\lambda_1 + x^*w, \lambda_2, \dots, \lambda_n$

Corollario

Gli autovalori di $\widetilde{G}$ sono $1, \widetilde{\lambda}_2, \dots, \widetilde{\lambda}_n$ con $|\lambda_i| < c$; in particolare $\gamma < c$

Domanda ingenua: perché non scegliamo $c \approx 0$?

Torniamo al pagerank

La matrice $\widetilde{G}$

- è positiva e quindi ha un **unico** autovettore positivo normalizzato;
- il metodo delle potenze converge con un parametro più piccolo di 0.85, il numero di passi richiesto è **moderato**

Che dire del costo computazionale di ogni passo?

Ad ogni passo un prodotto matrice-vettore è richiesto, $O(N^2)$ ops ad ogni passo

Troppo poiché N è enorme! La matrice G è troppo grande per essere tenuta in memoria.

Matrice sparsa

G è una matrice enorme, ma **sparsa**

Pochi link escono da ogni pagina $\rightarrow$ **pochi elementi non nulli** su ogni riga

- Come memorizzare una matrice sparsa?
- Come si calcola in modo efficiente il prodotto matrice sparsa-vettore?

Un problema in informatica

Le matrici sono di solito memorizzate come **array**

Una struttura migliore è la **lista**: ogni record contiene gli indici e gli elementi non nulli

- linked list
- array

la lista a puntatori è preferibile poiché la matrice G cambia ed è facile più facile aggiornare una lista a puntatori

Esempio

$$A = \begin{bmatrix} 2 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 5 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 \\ -3 & 0 & 1 & 4 & 0 \end{bmatrix}$$

$$\begin{aligned} &(1,1;2) \to (2,4;5) \to (4,1;4) \to (5,1;-3) \\ &\qquad \to (1,3;-1) \to (3,2;1) \to (5,4;4) \to (5,3;1) \end{aligned}$$

Esercizio

Descrivere un algoritmo che calcoli il prodotto di una matrice sparsa per un vettore con non più di $2k$ ops, dove k è il numero di elementi non nulli della matrice

Commento finale

Sì! G è sparsa, ma ci avevi detto che il calcolo viene fatto con $\widetilde{G} = cG + (1-c)ev^T$...

Commento finale

Sì! G è sparsa, ma ci avevi detto che il calcolo viene fatto con $\widetilde{G} = cG + (1-c)ev^T$...

Nessun problema!

$$\widetilde{G}w = cGw + (1-c)ev^T w = c(Gw) + (1-c)(v^T w)e,$$

si può calcolare Gw come sopra e $v^T w$ con un costo di $O(N)$ ops (assumiamo che $k \approx N$)

Calcolo del pagerank

Si parte da un vettore iniziale, per esempio $v_{i+1} = G^T v_i$.

$$v_1 = \begin{bmatrix} 0.2000 \\ 0.2000 \\ 0.2000 \\ 0.2000 \\ 0.2000 \end{bmatrix}$$

Calcolo del pagerank

Si parte da un vettore iniziale, per esempio $v_{i+1} = G^T v_i$.

$$v_2 = \begin{bmatrix} 0.3000 \\ 0.1167 \\ 0.2333 \\ 0.1667 \\ 0.1833 \end{bmatrix}$$

Calcolo del pagerank

Si parte da un vettore iniziale, per esempio $v_{i+1} = G^T v_i$.

$$v_3 = \begin{bmatrix} 0.2889 \\ 0.1139 \\ 0.2500 \\ 0.1889 \\ 0.1583 \end{bmatrix}$$

Calcolo del pagerank

Si parte da un vettore iniziale, per esempio $v_{i+1} = G^T v_i$.

$$v_4 = \begin{bmatrix} 0.2884 \\ 0.1102 \\ 0.2505 \\ 0.1977 \\ 0.1532 \end{bmatrix}$$

Calcolo del pagerank

Si parte da un vettore iniziale, per esempio $v_{i+1} = G^T v_i$.

$$v_5 = \begin{bmatrix} 0.2884 \\ 0.1088 \\ 0.2517 \\ 0.2006 \\ 0.1504 \end{bmatrix}$$

Calcolo del pagerank

Si parte da un vettore iniziale, per esempio $v_{i+1} = G^T v_i$.

$$v_6 = \begin{bmatrix} 0.2883 \\ 0.1084 \\ 0.2520 \\ 0.2019 \\ 0.1493 \end{bmatrix}$$

Calcolo del pagerank

Si parte da un vettore iniziale, per esempio $v_{i+1} = G^T v_i$.

$$v_7 = \begin{bmatrix} 0.2883 \\ 0.1082 \\ 0.2522 \\ 0.2024 \\ 0.1489 \end{bmatrix}$$

Calcolo del pagerank

Si parte da un vettore iniziale, per esempio $v_{i+1} = G^T v_i$.

$$v_8 = \begin{bmatrix} 0.2883 \\ 0.1081 \\ 0.2522 \\ 0.2026 \\ 0.1488 \end{bmatrix}$$

Calcolo del pagerank

Si parte da un vettore iniziale, per esempio $v_{i+1} = G^T v_i$.

$$v_9 = \begin{bmatrix} 0.2883 \\ 0.1081 \\ 0.2522 \\ 0.2027 \\ 0.1487 \end{bmatrix}$$

Calcolo del pagerank

Si parte da un vettore iniziale, per esempio $v_{i+1} = G^T v_i$.

$$v_{10} = \begin{bmatrix} 0.2883 \\ 0.1081 \\ 0.2522 \\ 0.2027 \\ 0.1487 \end{bmatrix}$$

Calcolo del pagerank

Si parte da un vettore iniziale, per esempio $v_{i+1} = G^T v_i$.

$$v_{11} = \begin{bmatrix} 0.2883 \\ 0.1081 \\ 0.2523 \\ 0.2027 \\ 0.1487 \end{bmatrix}$$

Calcolo del pagerank

Si parte da un vettore iniziale, per esempio $v_{i+1} = G^T v_i$.

$$v_{12} = \begin{bmatrix} 0.2883 \\ 0.1081 \\ 0.2523 \\ 0.2027 \\ 0.1487 \end{bmatrix}$$

Calcolo del pagerank

Si parte da un vettore iniziale, per esempio $v_{i+1} = G^T v_i$.

$$v_{13} = \begin{bmatrix} 0.2883 \\ 0.1081 \\ 0.2523 \\ 0.2027 \\ 0.1486 \end{bmatrix}$$

Calcolo del pagerank

Si parte da un vettore iniziale, per esempio $v_{i+1} = G^T v_i$.

$$v_{14} = \begin{bmatrix} 0.2883 \\ 0.1081 \\ 0.2523 \\ 0.2027 \\ 0.1486 \end{bmatrix}$$

Calcolo del pagerank

Si parte da un vettore iniziale, per esempio $v_{i+1} = G^T v_i$.

$$v_{15} = \begin{bmatrix} 0.2883 \\ 0.1081 \\ 0.2523 \\ 0.2027 \\ 0.1486 \end{bmatrix}$$

Calcolo del pagerank

Si parte da un vettore iniziale, per esempio $v_{i+1} = G^T v_i$.

$$v_{16} = \begin{bmatrix} 0.2883 \\ 0.1081 \\ 0.2523 \\ 0.2027 \\ 0.1486 \end{bmatrix}$$

Calcolo del pagerank

Si parte da un vettore iniziale, per esempio $v_{i+1} = G^T v_i$.

$$v_{17} = \begin{bmatrix} 0.2883 \\ 0.1081 \\ 0.2523 \\ 0.2027 \\ 0.1486 \end{bmatrix}$$

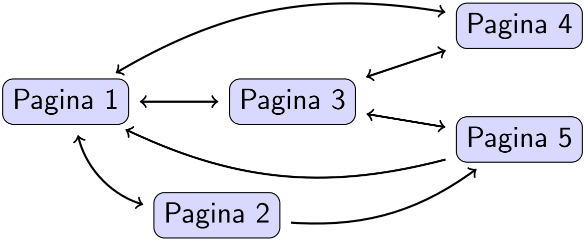
Calcolo del pagerank

Si parte da un vettore iniziale, per esempio $v_{i+1} = G^T v_i$.

$$v_{18} = \begin{bmatrix} 0.2883 \\ 0.1081 \\ 0.2523 \\ 0.2027 \\ 0.1486 \end{bmatrix}$$

Calcolo del pagerank

Si parte da un vettore iniziale, per esempio $v_{i+1} = G^T v_i$.



$$v_{18} = \begin{bmatrix} 0.2883 \\ 0.1081 \\ 0.2523 \\ 0.2027 \\ 0.1486 \end{bmatrix}$$

Complex networks

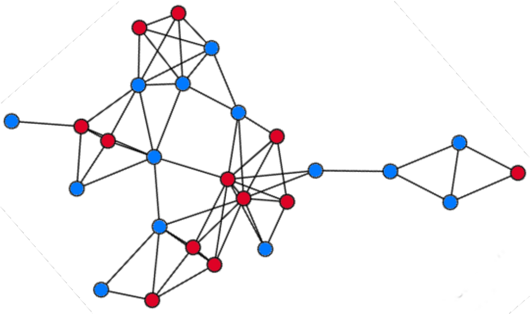
La discussione fatta per le pagine web può essere estesa a ogni problema che si modellizza con un grafo (diretto).

Un grafo è detto **complex network** se non mostra regolarità apparente.

Complex networks

Esempi di reti complesse si hanno in

- Informatica: Internet, World Wide Web, Social Networks
- Biologia: Food Webs, reti neurali, interazione di proteine
- Economia: World Trade
- Sociologia: **reti sociali**, reti di tossicodipendenti



Complex networks

Esempi di reti complesse si hanno in

- Informatica: Internet, World Wide Web, Social Networks
- Biologia: Food Webs, reti neurali, interazione di proteine
- Economia: World Trade
- Sociologia: reti sociali, reti di tossicodipendenti
- Accademia: citazioni di articoli, **collaborazioni scientifiche**

Search MSC | **Collaboration Distance** | Current Journals | Current Publications

MR Erdos Number = 4

Bruno Iannazzo	coauthored with	Raf Vandebril	MR3177955
Raf Vandebril	coauthored with	Gene Howard Golub	MR2191201
Gene Howard Golub	coauthored with	Alan J. Hoffman	MR0882452 (88f:41039)
Alan J. Hoffman	coauthored with	Paul Erdős[1]	MR0565328 (81b:05061)

Change First Author **Change Second Author** **New Search**

Free Tool Help | Support Mail

AMS American Mathematical Society

Mirror Sites Providence, RI USA

Complex networks

Esempi di reti complesse si hanno in

- Informatica: Internet, World Wide Web, Social Networks
- Biologia: Food Webs, reti neurali, interazione di proteine
- Economia: World Trade
- Sociologia: reti sociali, reti di tossicodipendenti
- Accademia: citazioni di articoli, collaborazioni scientifiche

solo per citarne qualcuna

Centralità

Il pagerank di Google è una soluzione del problema dell'importanza per le pagine web

Questo viene generalizzato ai complex network come **centralità**

- Il pagerank è l'**unico modello di importanza** per le pagine web?
- Ci sono **indici di centralità differenti** utili nelle applicazioni?

Hubs e autorità

Un altro ranking possibile della pagine web (facile da calcolare) si basa sulla classificazione in

- **autorità**: pagine con contenuti informativi;
- **hubs**: pagine che linkano qualcos'altro

Ogni pagina può essere un po' l'uno un po' l'altro.

Assegniamo a ogni pagina due valori: il **peso come autorità** $\{x_i\}_{i=1,\dots,N}$ e il **peso come hub** $\{y_i\}_{i=1,\dots,N}$.

Algoritmo HITS

La formulazione originale è stata data in termini dell'algoritmo **HITS** (Hypetext Induced Topics Search)

Si parte da $x^{(0)}$ e $y^{(0)}$, e si ottengono le sequenze

$$x_k^{(n)} = \sum_{j \in \mathcal{I}_k} y_j^{(n-1)}, \quad y_k^{(n)} = \sum_{j \in \mathcal{J}_k} x_j^{(n-1)}, \quad n = 1, 2, \ldots$$

normalizzate opportunamente.

$\mathcal{I}_k$ è l'insieme delle pagine che linkano k,
$\mathcal{J}_k$ è l'insieme delle pagine linkate da k.

Se una pagina è linkata da molte buone autorità allora è un buon hub;
se una pagina è linkata da molti buoni hub allora è una buona autorità.

Algoritmo HITS

Si può mostrare che l'algoritmo HITS è una variante del metodo delle potenze visto sopra applicato a delle matrici ottenute da A.

L'algoritmo HITS fornisce una **differente nozione di importanza** che dipende da cosa stiamo cercando

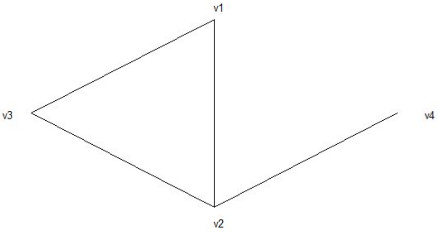
Nel World Wide Web, il pagerank di Google's **ha vinto** la gara, ma l'algoritmo HITS può essere utile in altri complex networks.

Indici di centralità

Alcune reti sono modellizzate come grafi non diretti (non si considerano link ma legami reciproci), ad esempio i social networks $\rightarrow$ la matrice di adiacenza è **simmetrica**.

Esempio:

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \qquad A^T = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} = A.$$



Indici di centralità

Indice più semplice: **il grado di centralità** del nodo $i \leftrightarrow$ è il numero di nodi ad esso connessi

Il grado di centralità rappresenta l'influenza immediata, ma è una misura molto grezza.

In epidemiologia si studia come un'infezione si diffonde in una rete

Un nodo infetto i può infettare i nodi a esso connessi

Il **rischio di un'epidemia** non è collegato solo a quanti nodi sono connessi con l'infetto i, ma anche a quanti nodi sono connessi con i nodi connessi con i e così via...

Indici di centralità

Per definire concetti più sofisticati di centralità consideriamo **potenze della matrice di adiacenza**

L'elemento i, j di A^2 è

$$(A^2)_{ij} = \sum_{k=1}^{N} a_{ik} a_{kj}$$

L'addendo k ($a_{ik} a_{kj}$) non è zero se $a_{ik} = a_{kj} = 1$

- $(A^2)_{ij} \neq 0$ se e solo se **c'è un percorso** di lunghezza 2 che unisce i e j;
- $(A^2)_{ij}$ conta il **numero di cammini** di lunghezza 2 che connettono i e j

Indici di centralità

L'elemento i, j della matrice A^n è

$$(A^n)_{ij} = \sum_{k_1, k_2, \ldots, k_{n-1}=1}^{N} a_{ik_1} a_{k_1 k_2} \cdots a_{k_{n-1} j}$$

L'addendo k non è nullo se $a_{ik_1} = a_{k_1 k_2} = \cdots = a_{k_{n-1} j} = 1$

- $(A^n)_{ij} \neq 0$ se e solo se **c'è un cammino** di lunghezza n che unisce i e j;
- $(A^n)_{ij}$ conta il **numero di cammini** di lunghezza n che connettono i e j

Indici di centralità

Possiamo immaginare che un nodo sia centrale se è **parte di molti cammini** che partono e finiscono con lui

Questa nozione è detta anche **subgraph centrality** perché conta (in qualche senso) a quanti mini-grafi il nodo appartiene.

Naturalmente, vogliamo dare **meno peso ai cammini più lunghi**

L'indice di Estrada

Sommando le potenze della matrice di adiacenza A in questo modo

$$\gamma_0 I + \gamma_1 A + \gamma_2 A^2 + \cdots + \gamma_n A^n + \cdots$$

otteniamo quella che si chiama una serie di potenze di A (somma infinita)

Se la somma è finita abbiamo in pratica definito una funzione $f(A)$ che è ancora una matrice il cui elemento $(f(A))_{ii}$ è la subgraph centrality del nodo i

L'indice di Estrada

La serie di potenze **più semplice** è l'esponenziale

$$\exp(A) = I + A + \frac{A^2}{2} + \frac{A^3}{3!} + \cdots + \frac{A^n}{n!} + \cdots$$

quindi otteniamo che la subgraph centrality del nodo i è $(\exp(A))_{ii}$, mentre la **connettività** tra i nodi i e j è $(\exp(A))_{ij}$

L'**indice di Estrada** è definito come la somma delle subgraph centrality $\sum_i (\exp(A))_{ii} = \text{trace}(\exp(A))$ e misura la connettività complessiva di un grafo