Similarità di sottografi nelle reti complesse

Gaspare Ferraro

Relatori
prof. Roberto Grossi
prof. Andrea Marino

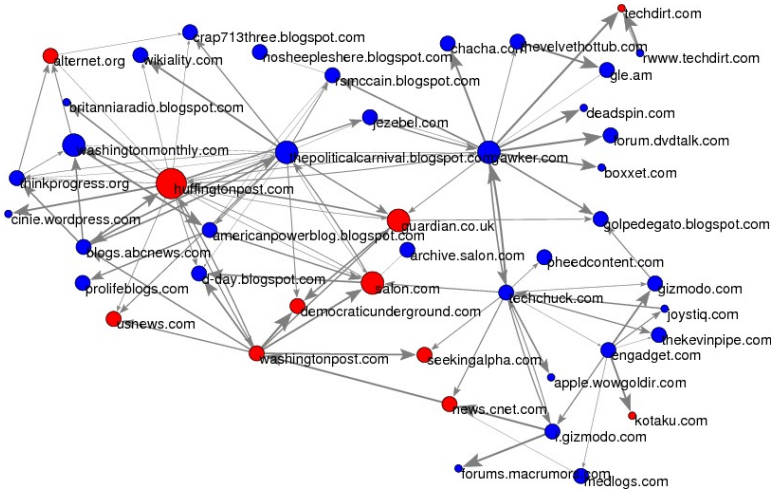
Università di Pisa
Dipartimento di Informatica

Pisa, 1 dicembre 2017

Il problema

Reti complesse

Grafi con caratteristiche topologiche non banali che occorrono modellando sistemi reali (quali social network, reti neurali, computer network).



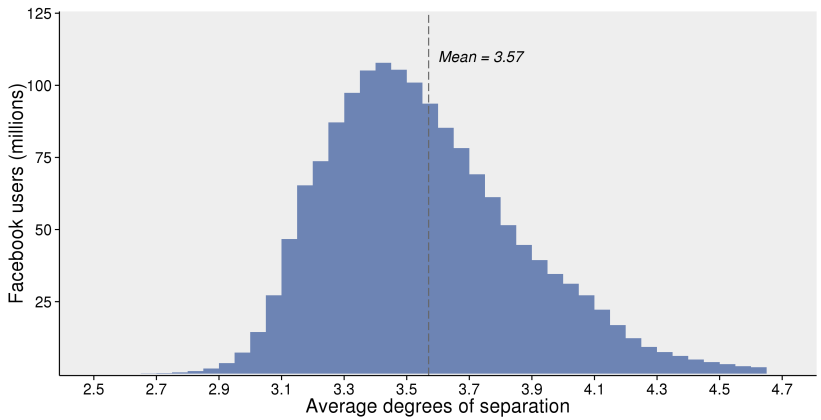
Diffusione delle notizie tra i vari siti e blog di informazione statunitensi
Fonte: SNAP Stanford

Rotte dei voli commerciali
Fonte: Bio Diaspora, Toronto

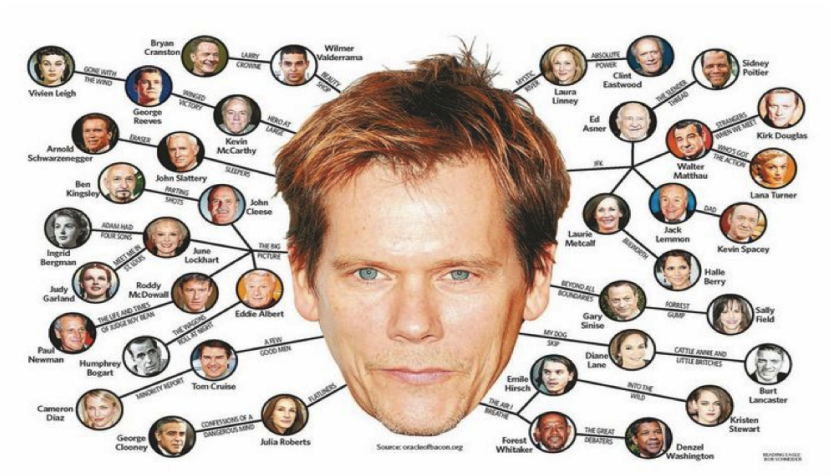
Sei gradi di separazione

"Ho letto che ognuno di noi su questo pianeta è separato dagli altri solo da sei persone. Sei gradi di separazione tra noi e tutti gli altri su questo pianeta [...] una tortura cinese essere così vicini ma dover trovare sei persone giuste per il collegamento."

Ouisa Kittredge, Six Degrees of Separation



In facebook la separazione media tra gli 1.6 miliardi di utenti registrati è 3.57.
Fonte: facebook research, Feb 2016



La distanza media di collaborazioni dall'attore Kevin Bacon è 3, il 98% degli attori è a distanza minore uguale a 6.
Fonte: IMDb, Ott 2017

n-grammi e reti etichettate

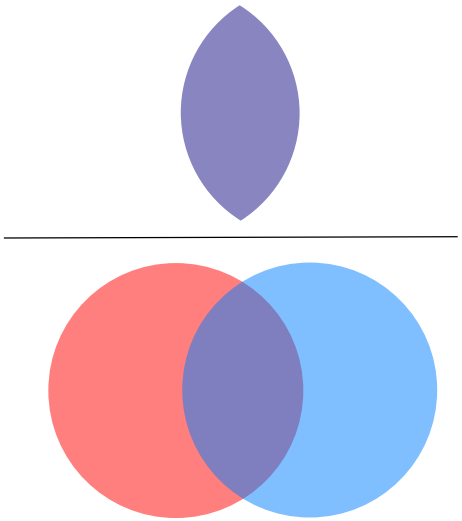
"Nessun uomo è un'isola, completo in se stesso; ogni uomo è un pezzo del continente, una parte del tutto."

John Donne

Indici di similarità

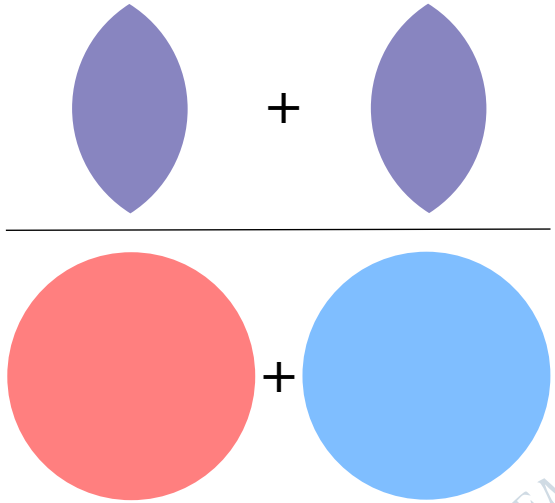
Jaccard

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$



Bray-Curtis

$$BC(A,B) = \frac{2 \times |A \cap B|}{|A| + |B|}$$

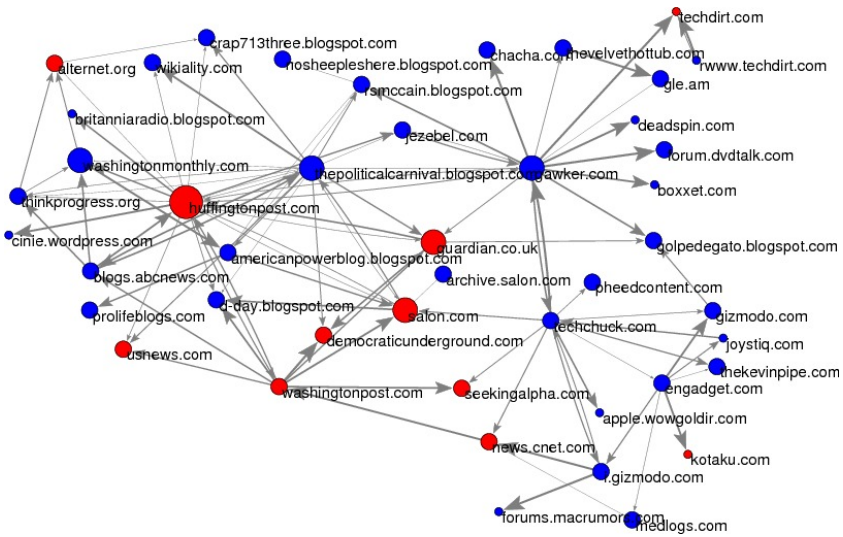


$$J(A,B) = BC(A,B) = 0 \text{ se } A \cap B = \emptyset$$

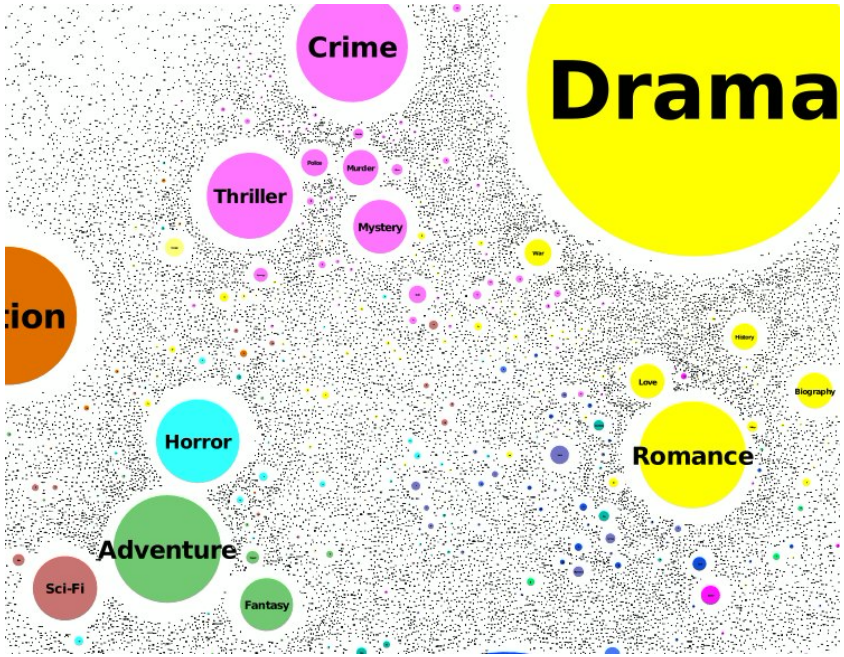
$$J(A,B) = BC(A,B) = 1 \text{ se } A = B$$

Il problema

Applicazioni pratiche



Diffusione delle notizie tra i vari blog e siti di informazione statunitensi
Fonte: SNAP Stanford



Interazione tra i film con attori in comune
Fonte: IMDb

Approcci di risoluzione

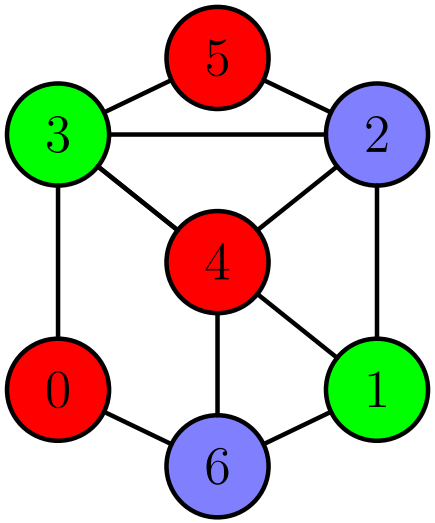
Ricerca esaustiva

Elenco tutti i possibili percorsi
Complessità

- Tempo: $O(|V|^q) \rightarrow$ Color Coding $\rightarrow O(2^q\ |V|)$
- Spazio: $O(|\Sigma|^q\ q) \rightarrow$ Sampling $\rightarrow O(rq)$

Color Coding

Coloro casualmente il grafo con q colori e mi limito ai path colorful (percorsi con colori non ripetuti)



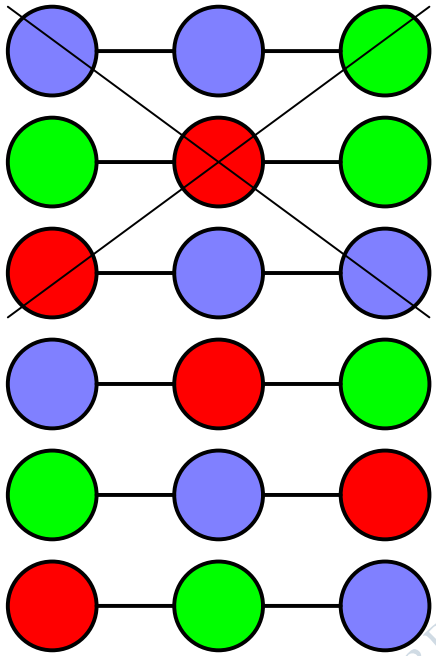
Il numero dei path è esponenzialmente ridotto di un fattore $q!/q^q \simeq e^{-q}$

Per $q = 3$ solo il $\sim 22.22\%$

Per $q = 6$ solo il $\sim 1.5\%$

$q!$ colorazioni accettabili

q^q possibili colorazioni



Esempi di possibili path

Sampling

F-Count

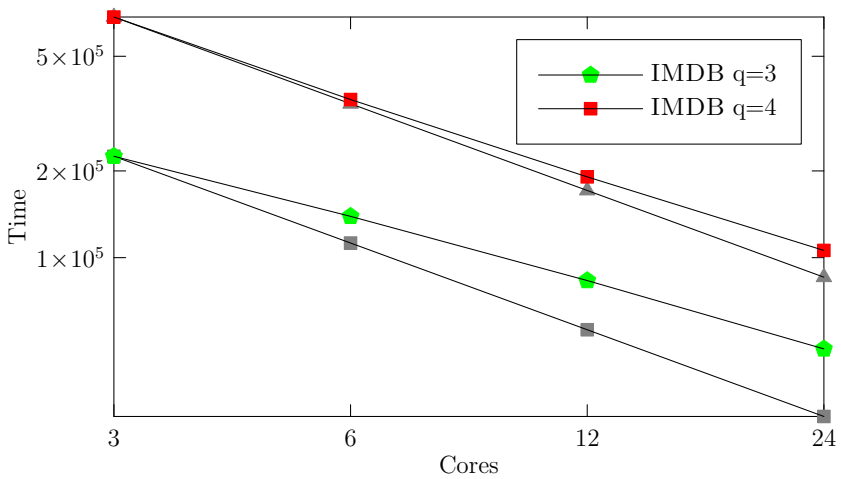
F-Samp

Risultati pratici

Color Coding

Tempi di esecuzione e memoria occupata

Dataset	q	Tempo	Memoria
NetInf	13	0.39s	11.20MiB
NetInf	14	0.81s	22.63MiB
NetInf	15	1.66s	45.21MiB
NetInf	16	3.47s	90.93MiB
IMDb	3	48.22s	17.94MiB
IMDb	4	105.94s	34.91MiB
IMDb	5	241.22s	69.01MiB
IMDb	6	557.48s	137.26MiB



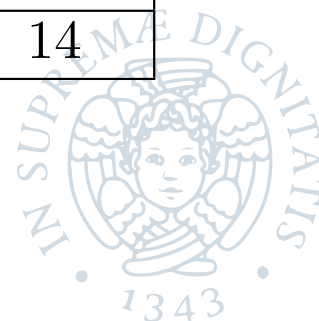
Scalabilità al variare dei cores usati

Query

					Tempi (in ms)		
DATASET	q	$\lvert A\rvert$	$\lvert B\rvert$	r	F-COUNT	F-SAMP	BASE
NETINF	3	100	100	1 000	20	4	2
NETINF	3	100	100	5 000	60	30	15
NETINF	5	100	100	1 000	2 682	426	3
NETINF	5	100	100	5 000	4 767	784	20
NETINF	7	100	100	100	5 455	4	2
NETINF	7	100	100	200	16 634	197	2
IMDB	3	10	10	100	5 035	66	1
IMDB	4	10	10	100	/	443	8
IMDB	5	10	10	100	/	781	12
IMDB	6	10	10	100	/	1 379	14

Tempi per il calcolo dell'indice di Bray-Curtis

r = Dimensione del campione



ϵ-approssimazione

Confronto a parità di livello di approssimazione ϵ

		F-COUNT			F-SAMP			BASE		
q	ϵ	r	T	VAR	r	T	VAR	r	T	VAR
3	0.20	2	1	0.0725	400	1	0.1194	420	1	0.1150
3	0.10	3	1	0.0692	1 000	1	0.0601	900	1	0.1338
3	0.05	4	1	0.0535	3 200	1	0.0273	1 500	1	0.1025
4	0.20	3	2	0.0677	1 300	1	0.1194	1 300	1	0.2424
4	0.10	5	4	0.0532	3 200	2	0.0992	2 500	2	0.1806
4	0.05	10	8	0.0518	8 000	4	0.0612	7 900	3	0.1081
5	0.20	5	6	0.0511	5 000	4	0.1678	6 000	3	0.2234
5	0.10	10	18	0.0370	20 000	12	0.0745	30 000	8	0.1234
5	0.05	20	58	0.0204	80 000	30	0.0376	/	/	/

Dati riferiti all'indice di Bray-Curtis su NETINF

Dimensione sottografi $|A| = |B| = 100$

$r =$ Dimensione del campione

$T =$ Tempo medio elaborazione (in millisecondi)

$VAR =$ Varianza indici

Nella pratica

Attore/Attrice	Attore/Attrice	BC index	FJ index
Stan Laurel	Oliver Hardy	0.936167	0.774053
Robert De Niro	Al Pacino	0.730935	0.231474
Woody Allen	Meryl Streep	0.556071	0.222857
Meryl Streep	Roberto Benigni	0.482909	0.160181

IMDb, Similarità tra ego-network di attori famosi (F-Samp)

Sito	Sito	BC index	FJ index
cnn.com	huffpost.com	0.936167	0.774053
nytimes.com	cnn.com	0.730935	0.231474
huffpost.com	nytimes.com	0.556071	0.222857

NetInf, Similarità tra siti di informazione (F-Samp)

Conclusioni

F-Count

Pro:

- Accurato anche con campioni di piccole dimensioni
- Varianza ridotta

Contro:

- Lento su grafi di elevate dimensioni
- Preprocessing grafo (una volta sola)

F-Samp

Pro:

- Efficiente anche in grafi di elevate dimensioni
- Varianza ridotta

Contro:

- Necessita di campioni di grandi dimensioni
- Preprocessing grafo (una volta sola)

Base

Pro:

- Efficiente anche in grafi di elevate dimensioni

Contro:

- Varianza elevata
- Necessita di campioni di grandi dimensioni
- Può non convergere al valore esatto

Fine

Grazie per l'attenzione