

# 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

# Parte I

## Il problema



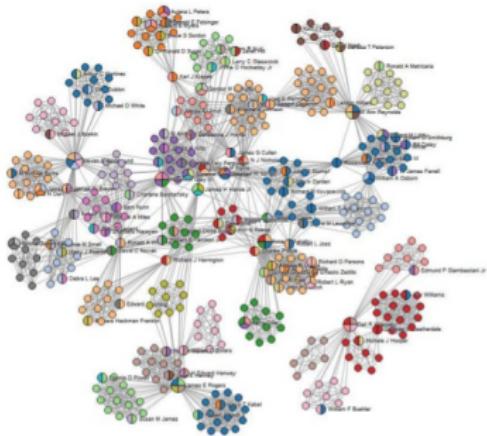
# Reti complesse

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



# Reti complesse

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

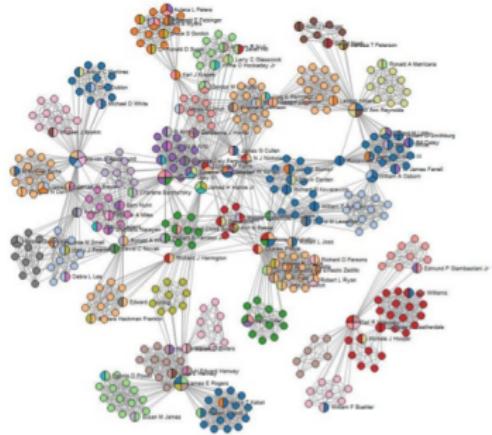


Cluster di amicizie in un social network



# Reti complesse

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



Cluster di amicizie in un social network



Rotte dei voli commerciali

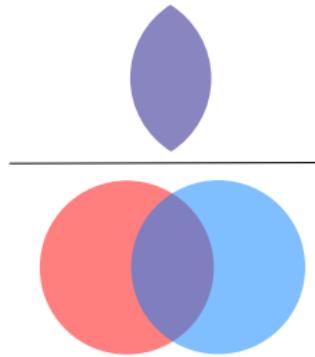
# Indici di similarità



# Indici di similarità

## Jaccard

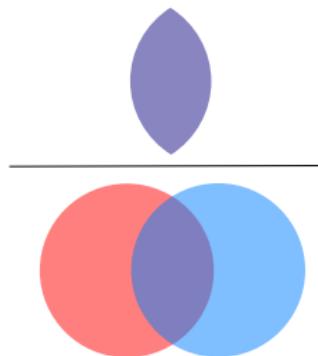
$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$



# Indici di similarità

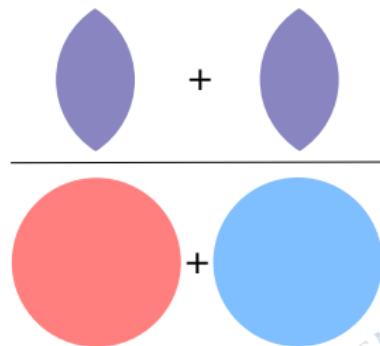
## Jaccard

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



## Bray-Curtis

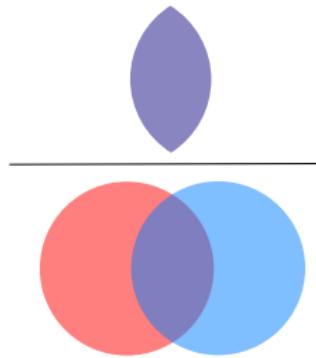
$$BC(A, B) = \frac{2 \times |A \cap B|}{|A| + |B|}$$



# Indici di similarità

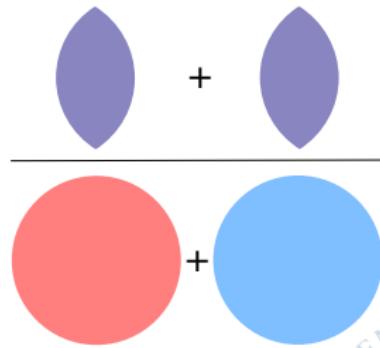
## Jaccard

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



## Bray-Curtis

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



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

$J(A, B) = BC(A, B) = 1$  se  $A = B$



# Reti etichettate e $q$ -grammi

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

*John Donne*

*Analizzare non il solo nodo, ma anche la sua interfaccia verso l'esterno!*



# Reti etichettate e $q$ -grammi

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

*John Donne*

*Analizzare non il solo nodo, ma anche la sua interfaccia verso l'esterno!*

Come modellare le interazioni?



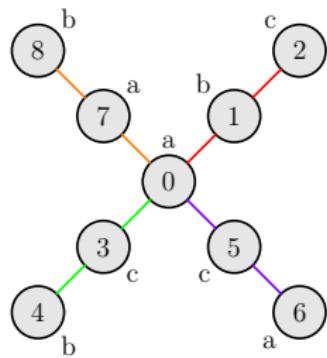
# Reti etichettate e $q$ -grammi

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

John Donne

*Analizzare non il solo nodo, ma anche la sua interfaccia verso l'esterno!*  
Come modellare le interazioni?

Rete etichettata



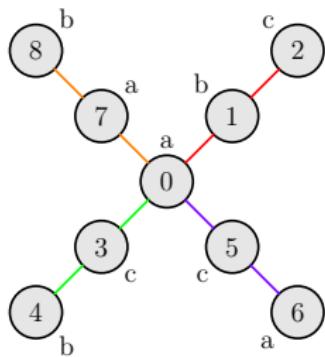
# Reti etichettate e $q$ -grammi

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

John Donne

*Analizzare non il solo nodo, ma anche la sua interfaccia verso l'esterno!*  
Come modellare le interazioni?

Rete etichettata



**$q$ -grammi:** sottosequenza di  $q$  elementi consecutivi in un testo

+

**$q$ -path:** cammino di  $q$  nodi *distinti* collegati in un grafo



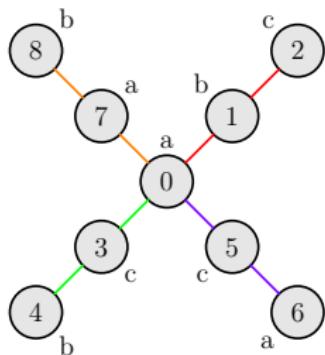
# Reti etichettate e $q$ -grammi

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

John Donne

Analizzare non il solo nodo, ma anche la sua interfaccia verso l'esterno!  
Come modellare le interazioni?

Rete etichettata



**$q$ -grammi:** sottosequenza di  $q$  elementi consecutivi in un testo

+

**$q$ -path:** cammino di  $q$  nodi *distinti* collegati in un grafo

Esempio 3-grammi che terminano in 0:

- $(2 - 1 - 0)$ : **cba**
- $(4 - 3 - 0)$ : **bca**
- $(6 - 5 - 0)$ : **aca**
- $(8 - 7 - 0)$ : **baa**



# Frequenze dei $q$ -grammi

## Notazione:

$f_X[w] = y \rightarrow$  Il  $q$ -gramma  $w$  ha frequenza  $y$  nei  $q$ -path che terminano in nodi di  $X$

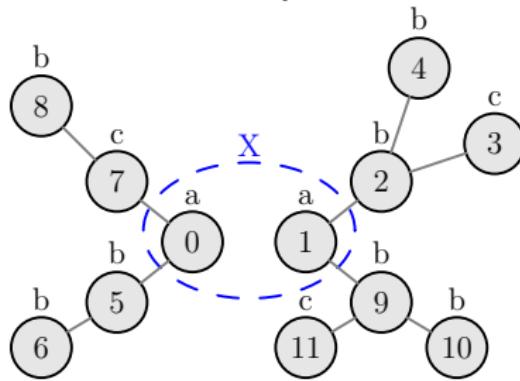


# Frequenze dei $q$ -grammi

## Notazione:

$f_X[w] = y \rightarrow$  Il  $q$ -gramma  $w$  ha frequenza  $y$  nei  $q$ -path che terminano in nodi di  $X$

## Esempio:



Dato  $X = \{0, 1\}$  e  $q = 3$  abbiamo:

- $f_X[bba] = 3$  (path: 4-2-1, 6-5-0, 10-9-1).
- $f_X[bca] = 1$  (path: 8-7-0).
- $f_X[cba] = 2$  (path: 3-2-1, 11-9-1).



## Il problema

Dato un grafo  $G = (V, E, L)$ , etichettato su un alfabeto  $\Sigma$ , ed un intero  $q$ , calcolare la similarità tra due porzioni di grafo  $A, B \subset V$  in base alle frequenze dei  $q$ -grammi dei  $q$ -path che terminano in nodi di  $A$  e  $B$ .



## Il problema

Dato un grafo  $G = (V, E, L)$ , etichettato su un alfabeto  $\Sigma$ , ed un intero  $q$ , calcolare la similarità tra due porzioni di grafo  $A, B \subset V$  in base alle frequenze dei  $q$ -grammi dei  $q$ -path che terminano in nodi di  $A$  e  $B$ .

Estendiamo i due indici ai  $q$ -grammi:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \implies J(A, B) = \frac{\sum_{x \in \Sigma^q} \min(f_A[x], f_B[x])}{\sum_{x \in \Sigma^q} f_{A \cup B}[x]}$$

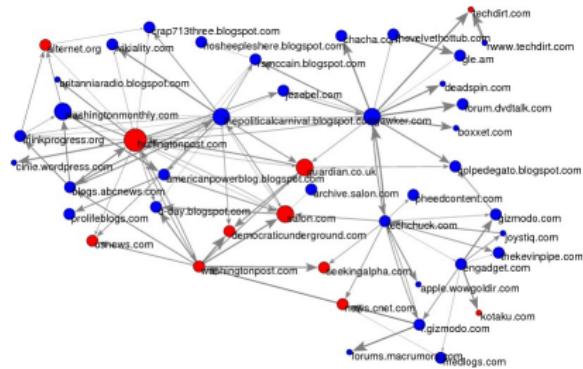
$$BC(A, B) = \frac{2 \times |A \cap B|}{|A| + |B|} \implies BC(A, B) = \frac{2 \times \sum_{x \in \Sigma^q} \min(f_A[x], f_B[x])}{\sum_{x \in \Sigma^q} (f_A[x] + f_B[x])}$$

# Applicazioni pratiche



# Applicazioni pratiche

## NetInf

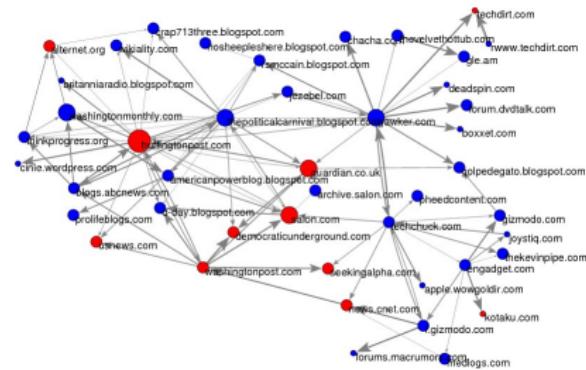


Diffusione delle notizie tra i vari blog e siti di informazione statunitensi  
Fonte: *SNAP Stanford*



# Applicazioni pratiche

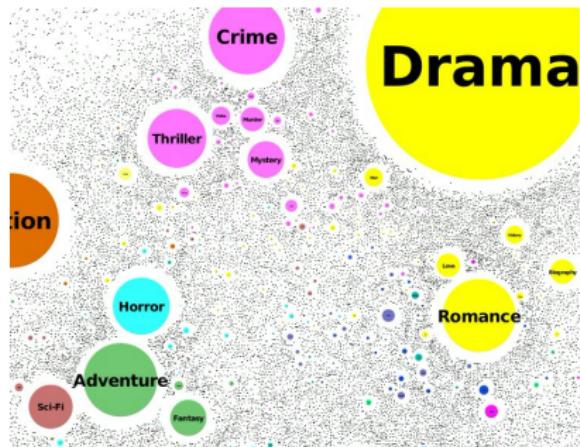
NetInf



Diffusione delle notizie tra i vari blog e siti di informazione statunitensi

Fonte: SNAP Stanford

IMDb



Interazione tra i film con attori in comune

Fonte: IMDb

## Parte II

# Approcci di risoluzione



# Ricerca esaustiva

Elenco tutti i possibili percorsi

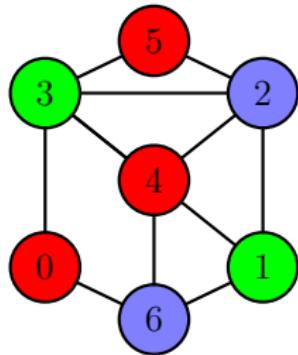
Complessità

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



# Color Coding

Coloriamo casualmente il grafo con  $q$  colori e ci limitiamo ai path colorful (percorsi con colori non ripetuti)

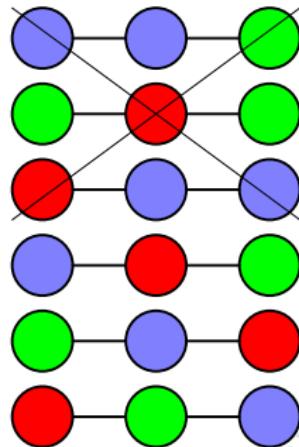


Il numero dei path è esponenzialmente ridotto di un fattore  $q!/q^q \simeq e^{-q}$

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



# Parte III

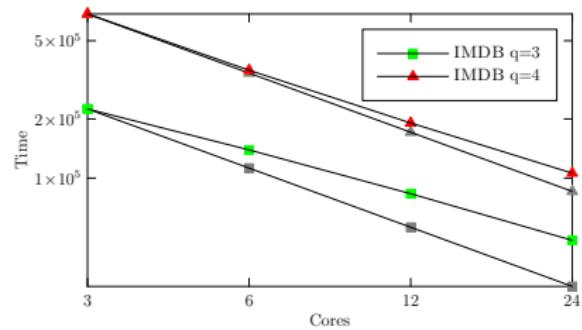
## Risultati pratici



## preprocessing: 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: F-Count / F-Samp / Base

DATASET	$q$	$ A $	$ B $	$r$	Tempi (in ms)		
					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



# $\epsilon$ -approssimazione

## Confronto a parità di livello di approssimazione $\epsilon$

		F-COUNT			F-SAMP			BASE		
$q$	$\epsilon$	$r$	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
nytimes.com	huffpost.com	0.760524	0.388977
nytimes.com	washingtonpost.com	0.732766	0.366383
nytimes.com	sportingnews.com	0.330400	0.166200
nytimes.com	rollingstone.com	0.056660	0.034336

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



## 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."*

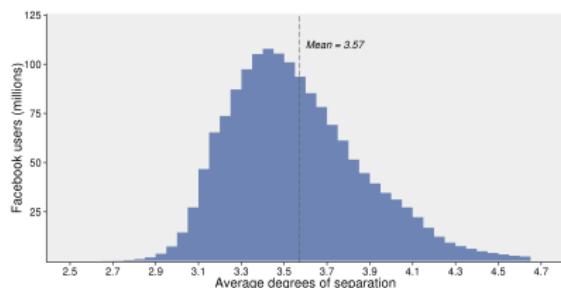
*Ouisa Kittredge, Six Degrees of Separation*



# 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.

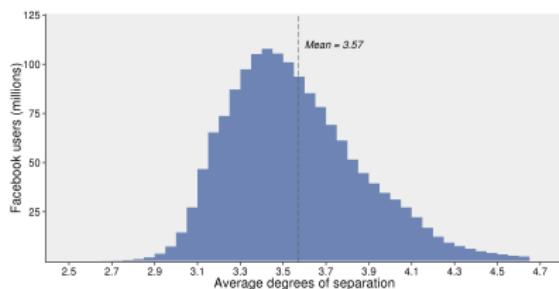
Fonte: Facebook Research, Feb 2016



# Sei gradi di separazione

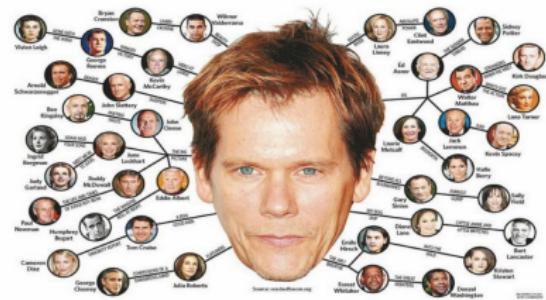
*"Ho letto che ognuno di noi su questo pianeta è separato dagli altri solo da sei persone. Sei gradi di separazione tra noi e tutti gli altri su questo pianeta [...] una tortura cinese essere così vicini ma dover trovare sei persone giuste per il collegamento."*

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 un attore, in termini di collaborazioni, da Kevin Bacon è 3, il 98% degli attori è a distanza  $\leq 6$ .

Fonte: IMDb, Ott 2017