Università degli Studi di Milano


# TECNICHE ONLINE PER L'INDIVIDUAZIONE DI SPAM IN UN WEB CRAWLER

Relatore

*Paolo Boldi*

Correlatore

*Sebastiano Vigna*

Candidato

*Antonio Luca*

# Obbiettivo della tesi

Obbiettivo di questa tesi è stata l'analisi delle varie tecniche di spam detection descritte in letteratura al fine di valutarne il comportamento e vagliare la possibilità di utilizzo di tali tecniche online.

# Obbiettivo della tesi

Obbiettivo di questa tesi è stata l'analisi delle varie tecniche di spam detection descritte in letteratura al fine di valutarne il comportamento e vagliare la possibilità di utilizzo di tali tecniche online.

## Struttura della tesi

- Classificazione delle varie tecniche di spam detection sulla base dei segnali utilizzati.
- Analisi degli algoritmi offline durante la fase di crawling.

# Motivazioni

- I risultati di tale lavoro sarranno utilizzati per essere integrati all'interno di un web crawler distribuito ad alte prestazioni per il futuro sviluppo di un modulo di spam detection.
- L'esigenza di tale modulo è sorta a seguito dello sviluppo, presso il Dipartimento di Informatica, di un crawler chiamato BUbiNG, altamente configurabile ma privo al momento di qualunque forma di rilevazione di siti e contenuti malevoli.

# Il fenomeno del web spam

## Le cause

- Col crescere delle dimensioni del web, aumenta la difficoltà di una pagina di comparire tra i primi risultati di un motore di ricerca per una data quey.

# Il fenomeno del web spam

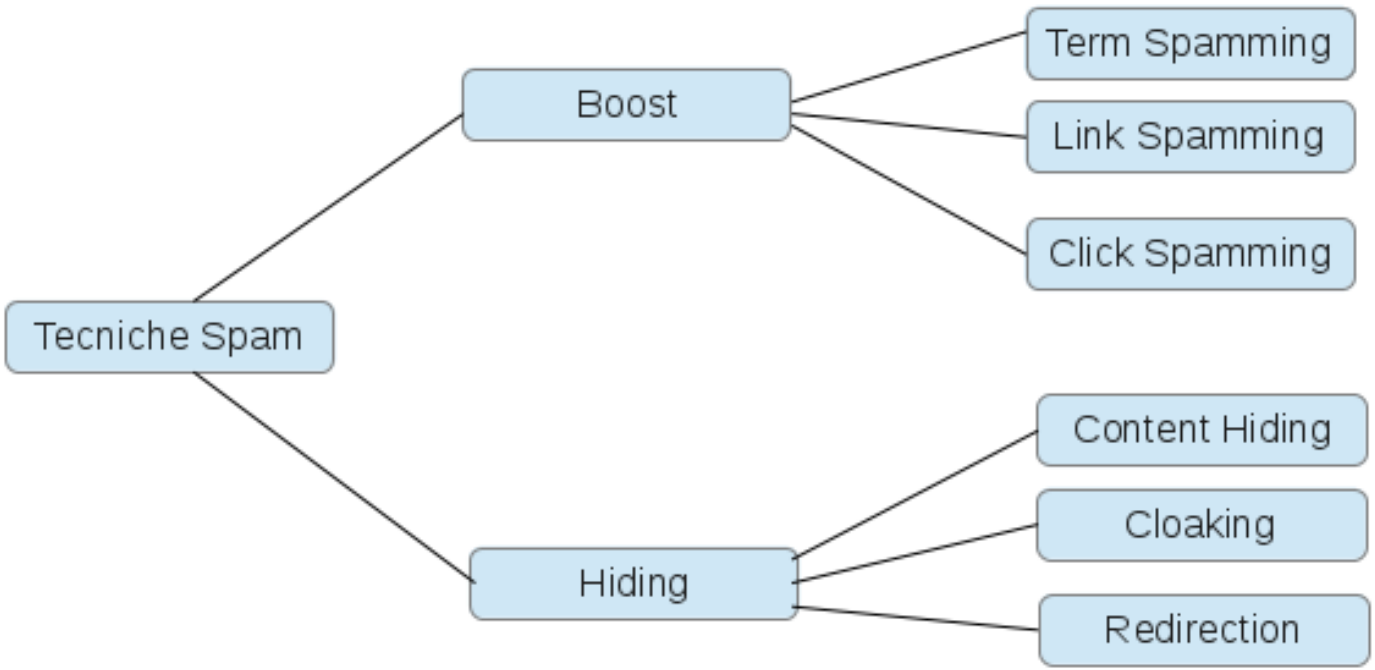
## Le cause

- Col crescere delle dimensioni del web, aumenta la difficoltà di una pagina di comparire tra i primi risultati di un motore di ricerca per una data quey.

## Conseguenze

- Sviluppo di meccanismi di spam per tentare di ingannare gli algoritmi dei motori di ricerca al fine di ottenere un rank maggiore per una data pagina web.

# Tecniche spam



# Tecniche di boost: term spamming

Il term spamming sfrutta alcuni punti di una pagina web per manipolarne il ranking.

# Tecniche di boost: term spamming

Il term spamming sfrutta alcuni punti di una pagina web per manipolarne il ranking.

## Tassonomia

- Body spam;
- Title spam;
- Meta tag spam;
- Anchor text spam;
- URL spam.

# Tecniche di boost: link spamming

Il link spamming è un tipo di spam che fa uso della struttura dei link tra le pagine web per favorire il rank di una pagina target $t$.

# Tecniche di boost: link spamming

Il link spamming è un tipo di spam che fa uso della struttura dei link tra le pagine web per favorire il rank di una pagina target $t$.

## Tassonomia

- Outgoing link spam;
- Incoming link spam (Honeypot, Blog, Forum, Wiki, Scambio di link, Domini scaduti, Spam farm).

# Click spamming

I motori di ricerca utilizzano dati sul flusso dei click per regolare le funzioni di ranking, quindi gli spammer generano clik fraudolenti per manipolare il comportamento di queste funzioni in modo tale da fare avere un rank migliore ai loro siti.

# Tecniche di hiding

## Tassonomia

- Content hiding;
- Cloaking;
- Redirection.

# Tecniche di spam detection

# Tecniche di spam detection

## Classificazione

# Tecniche di spam detection

## Classificazione

- Tecniche basate sul contenuto;

# Tecniche di spam detection

## Classificazione

- Tecniche basate sul contenuto;
- Tecniche basate sul grafo del web;

# Tecniche di spam detection

## Classificazione

- Tecniche basate sul contenuto;
- Tecniche basate sul grafo del web;
- Tecniche basate su segnali eterogenei.

# Tecniche basate sul contenuto - 1

Un metodo per identificare lo spam basandosi sul contenuto di una pagina web è quello di analizzare alcune feature delle pagine spam e confrontarle con le medesime feature delle pagine non spam al fine di ottenere dei valori con cui stimare la natura della pagina web.

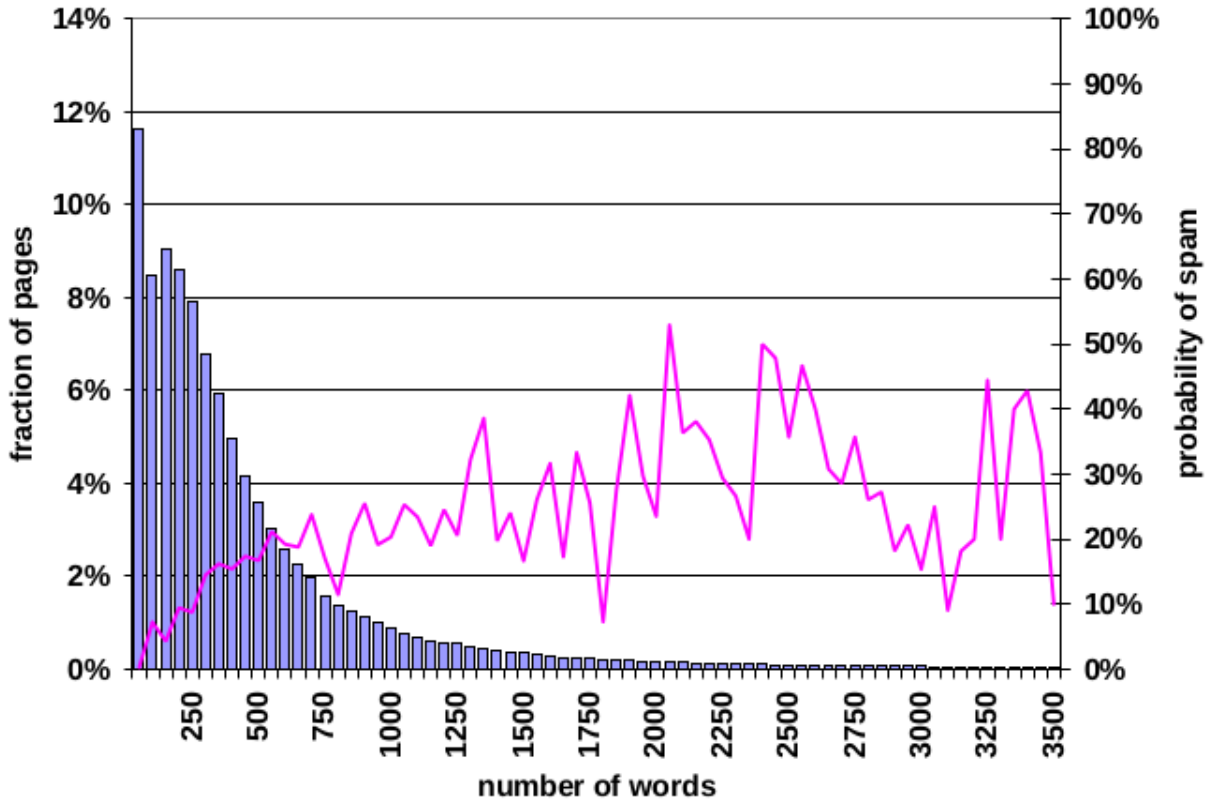
# Tecniche basate sul contenuto - 1

Un metodo per identificare lo spam basandosi sul contenuto di una pagina web è quello di analizzare alcune feature delle pagine spam e confrontarle con le medesime feature delle pagine non spam al fine di ottenere dei valori con cui stimare la natura della pagina web.

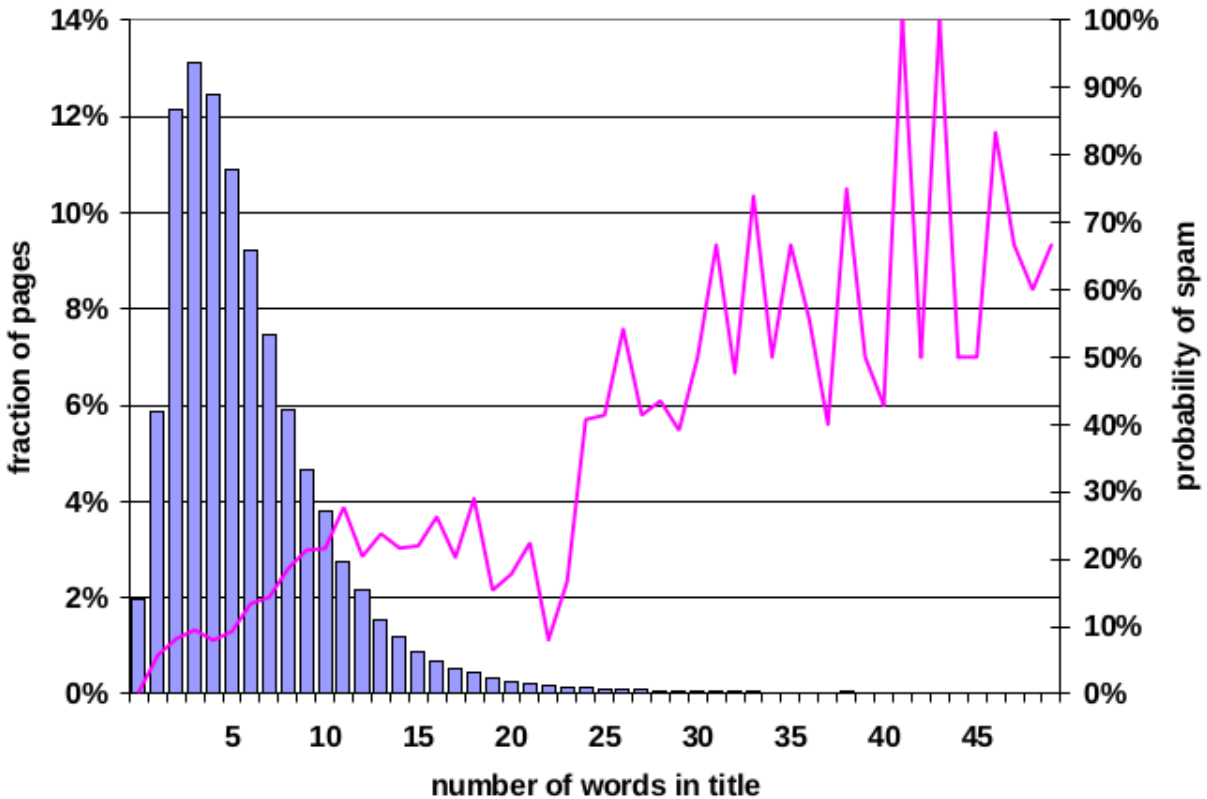
## Esempi di feature

- Numero di parole all'interno della pagina (Keyword Stuffing);
- Numero di parole all'interno dei titoli delle pagine;
- Lunghezza media delle parole all'interno delle pagine;
- Lunghezza testo all'interno del tag $< a >$;
- Frazione di contenuto visibile;

# Feature: Numero di parole nel body



# Feature: Numero di parole nel titolo



# Tecniche basate sul contenuto - 2

- Viene utilizzata la *Kullback-Leibler Divergence* (KLD) per misurare la divergenze tra le distribuzioni di probabilità dei termini di pagine web, applicandola a unità di testo della pagina di partenza e di quella linkata
- La Kullback-Leibler (KL) è una misura asimmetrica della divergenza che misura quanto male una distribuzione di probabilità $M_q$ riesce a modellare $M_d$

$$KLD(T_1||T_2) = \sum_{t \in T_1} P_{T_1}(t) \log \frac{P_{T_1}(t)}{P_{T_2}(t)} \quad (1)$$

# KLD: Tipo di sorgenti

# KLD: Tipo di sorgenti

## Pagina di Partenza

- testo delle ancore
- testo intorno alle ancore
- termini nell'URL

# KLD: Tipo di sorgenti

## Pagina di Partenza

- testo delle ancore
- testo intorno alle ancore
- termini nell'URL

## Pagina di Arrivo

- titolo della pagina
- contenuto della pagina
- meta tag

# KLD: Uso delle sorgenti

## Combinazione

- testo delle ancore - contenuto
- testo vicino alle ancore - contenuto
- termini nell'URL - contenuto
- testo delle ancore - titolo
- testo intorno alle ancore - titolo
- termini nell'URL - titolo
- titolo - contenuto
- metatag

# Tecniche basate sul grafo

## Tecniche base

- TrustRank
- Anti-trust Rank
- Spam mass

# Tecniche basate sul grafo

## Tecniche base

- TrustRank
- Anti-trust Rank
- Spam mass

Utilizzano una versione personalizzata di PageRank:

$$\alpha G + (1 - \alpha)1v^t \tag{2}$$

# TrustRank

TrustRank tenta di assegnare un valore di rank maggiore alle pagine non spam rispetto alle pagine spam partendo da un insieme di pagine non spam.

# TrustRank

TrustRank tenta di assegnare un valore di rank maggiore alle pagine non spam rispetto alle pagine spam partendo da un insieme di pagine non spam.

## Assunzione

- Per determinare le pagine non spam, viene fatta un'assunzione empirica chiamata isolazione approssimata dell'insieme delle pagine buone, la quale afferma che le pagine non spam raramente punteranno a delle pagine spam.
- Gli sviluppatori di pagine web non spam non hanno interesse nel linkare pagine spam (a meno che tramite l'uso di tecniche come l'honeypot vengano "ingannati").

# TrustRank

- TrustRank quindi è una versione personalizzata di PageRank dove il vettore di preferenza $v$ non rappresenta una distribuzione uniforme su tutte le pagine del grafo $G$ ma una distribzione personalizzata dalle pagine del seedset di partenza.

$$\alpha G + (1 - \alpha)1v^t$$

- I campi del vettore $v$ avranno valore 1 se corrisponderanno al nodo del seedset altrimenti 0.

# Anti-trust Rank

- Partendo dalla stessa intuizione dell'isolamento approsimato
- Anti-trust Rank utilizza un seed set iniziale composto da pagine spam
- Quindi una pagina spam (conosciuta) è linkata da un'altra pagina spam in quanto una pagina non spam non ne avrebbe il motivo
- Come per TrustRank, Anti-trust Rank viene calolato usando Pagerank personalizzato sul grafo trasposto (visto che adesso si prendono in consierazioni i link in entrata)
- Anti-trust Rank assegna un rank maggiore alle pagine spam

# Spam mass

- Per riconoscere una spam farm si parte dal presupposto che i nodi della spam farm hanno dei link uscenti verso delle pagine target $t$ per aumentarne il rank
- Spam mass è una misura che valuta l'impatto dello spam sul rank di una pagina
- Vengono assegnati due valori: PageRank e Spam mass

# Determinazione della spam farm

Partendo da un insieme $(\tilde{V})^{+}$ composto da pagine non spam conosiute

# Determinazione della spam farm

Partendo da un insieme $(\tilde{V})^+$ composto da pagine non spam conosiute

## Si calcolano due misure

- Spam mass assoluta: $\tilde{M}_x = p_x - p'_x$
- Spam mass relativa: $\tilde{m}_x = (p_x - p'_x)/p_x = 1 - p'_x/px$

dove $p$ è il *pagerank* dei nodi basato su una distribuzione uniforme mentre $p'$ è *pagerank* basato sull'insieme $(\tilde{V})^+$.
Viene usato un valore di soglia per determinare se una pagina è parte di una spam farm.

# Tecniche eterogenee: Header HTTP

- Questo metodo utilizza le informazioni racchiuse all'interno degli header HTTP per determinare le pagine spam
- Può essere usato come supporto ad altri metodi descritti in precedenza e può essere utilizzato in modo dinamico durante la fase di download delle pagine

# Tecniche eterogenee: Header HTTP

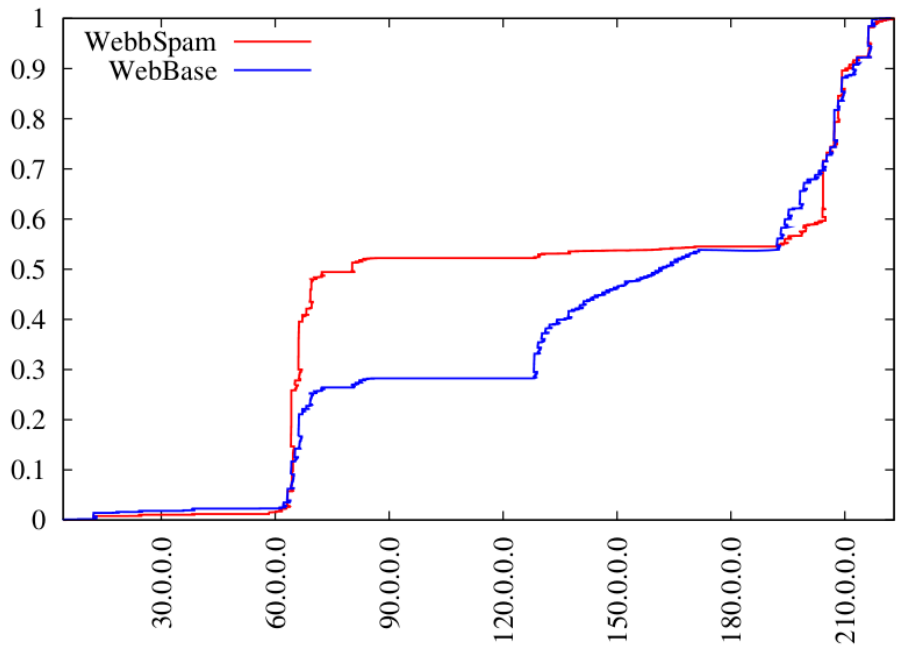
- Questo metodo utilizza le informazioni racchiuse all'interno degli header HTTP per determinare le pagine spam
- Può essere usato come supporto ad altri metodi descritti in precedenza e può essere utilizzato in modo dinamico durante la fase di download delle pagine

## Funzionamento

- Dopo aver effettuato la richiesta HTTP al server di una pagina vengono interpretati solo gli header della riposta HTTP
- Successivamente viene azionato un classificatore per valutare gli header come spam o non spam
- Se gli header vengono classificati come non spam allora si continua con la lettura del resto della pagina.

# Header HTTP - Esempio

Analizzando i valori dei campi degli header HTTP si nota che alcuni di essi sono più frequenti nelle pagine spam invece delle pagine non spam



# Test

- Si è scelto, di valutare l'efficacia di due algoritmi link based offline (TrustRank e Anti-trust Rank ) durante l'operazione di crawling ovvero in modo online
- Il razionale di tale analisi è stato la valutazione dell'operabilità di tali algoritmi durante l'esecuzione online e il confronto delle prestazioni rispetto all'utilizzo convenzionale offline

# Test

- Si è scelto, di valutare l'efficacia di due algoritmi link based offline (TrustRank e Anti-trust Rank ) durante l'operazione di crawling ovvero in modo online
- Il razionale di tale analisi è stato la valutazione dell'operabilità di tali algoritmi durante l'esecuzione online e il confronto delle prestazioni rispetto all'utilizzo convenzionale offline

**Nota**

La simulazione della fase di crawling stata fatta tramite una visita in ampizza su grafo.

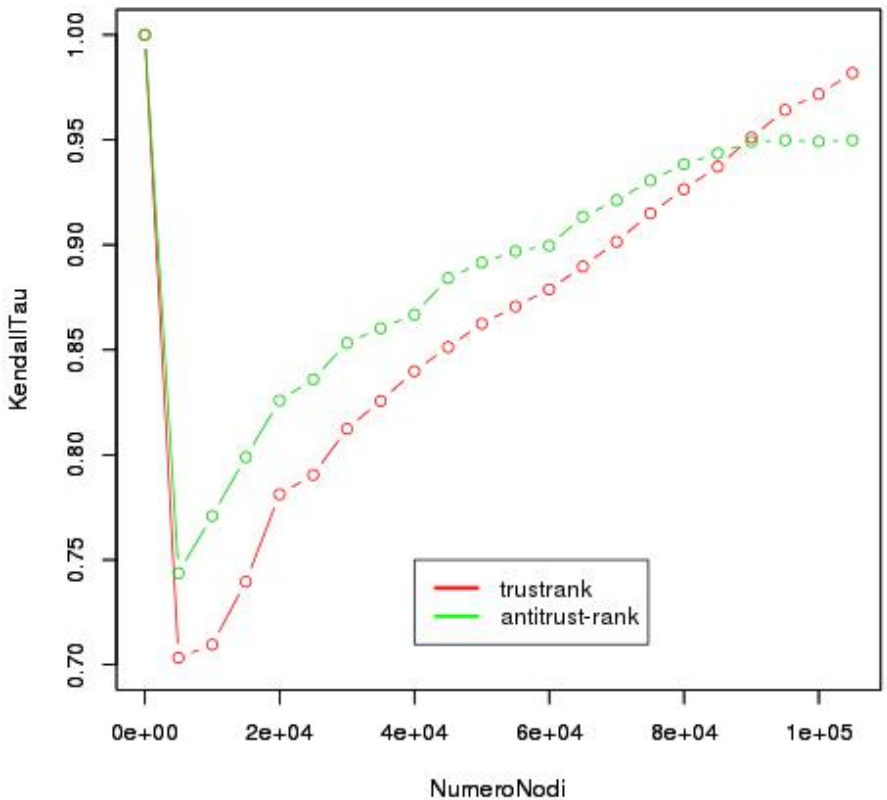
# Test 1

- Calcolo della distanza, attraverso l'utilizzo della Tau di Kendall $\tau_t$, tra il vettore $t$ di TrustRank ricavato sull'intero grafo $G$ e il vettore $t_i$ di TrustRank calcolato sul grafo temporaneo $G_v$ ricavato ad ogni intervallo di nodi visitati lungo una visita in ampiezza $v$ con nodo sorgente $s$.

# Test 1

- Calcolo della distanza, attraverso l'utilizzo della Tau di Kendall $\tau_t$, tra il vettore $t$ di TrustRank ricavato sull'intero grafo $G$ e il vettore $t_i$ di TrustRank calcolato sul grafo temporaneo $G_v$ ricavato ad ogni intervallo di nodi visitati lungo una visita in ampiezza $v$ con nodo sorgente $s$.
- Calcolo della distanza, attraverso l'utilizzo della Tau di Kendall $\tau_a$, tra il vettore $a$ di Anti-trust Rank ricavato sull'intero grafo $G$ e il vettore $a_i$ di Anti-trust Rank calcolato sul grafo temporaneo $G_v$ ricavato ad ogni intervallo di nodi visitati lungo una visita in ampiezza $v$ con nodo sorgente $s$.

# Test 1: Grafici



# Test 1: Risultati

- Il comportamento tra $\tau_t$ e $\tau_a$ indica che TrustRank online è meno efficace di Anti-trust Rank nell'approssimare il comportamento offline.

# Test 1: Risultati

- Il comportamento tra $\tau_t$ e $\tau_a$ indica che TrustRank online è meno efficace di Anti-trust Rank nell'approssimare il comportamento offline.
- Perciò tra questi due algoritmi quello che si adatta meglio nell'utilizzo in modalità online è Anti-trust rank perché tende ad approssimare fin da subito il comportamento offline.

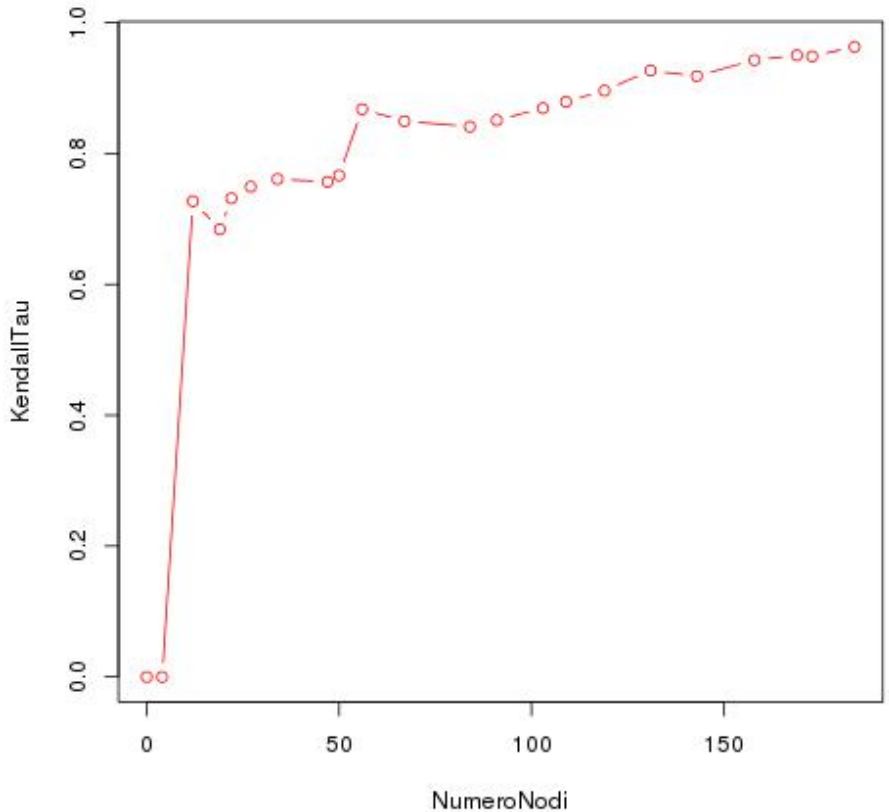
# Test 2

- Calcolo della distanza, attraverso l'utilizzo della Tau di Kendall $\tau_t$, tra il vettore $t$ di TrustRank ricavato sull'intero grafo $G$ e il vettore $t_i$ di TrustRank calcolato sul grafo temporaneo $G_v$, prendendo in considerazione i soli nodi spam.

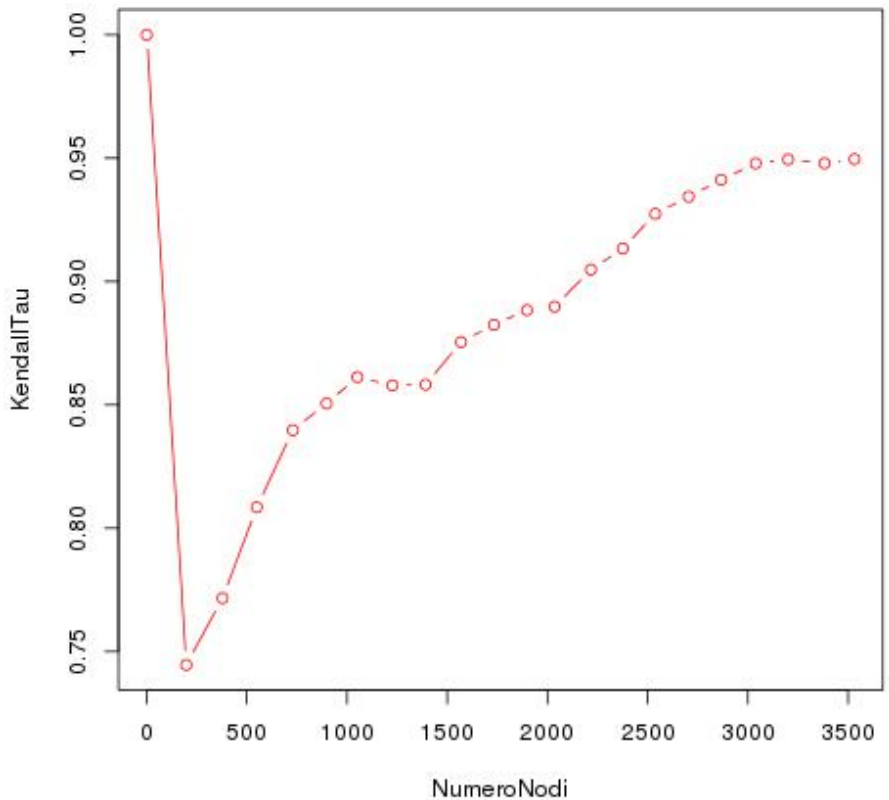
# Test 2

- Calcolo della distanza, attraverso l'utilizzo della Tau di Kendall $\tau_t$, tra il vettore $t$ di TrustRank ricavato sull'intero grafo $G$ e il vettore $t_i$ di TrustRank calcolato sul grafo temporaneo $G_v$, prendendo in considerazione i soli nodi spam.
- Calcolo della distanza, attraverso l'utilizzo della Tau di Kendall $\tau_a$, tra il vettore $a$ di Anti-trust Rank ricavato sull'intero grafo $G$ e il vettore $a_i$ di Anti-trust Rank calcolato sul grafo temporaneo $G_v$, prendendo in considerazione i soli nodi non spam.

# Test 2: Grafico TrustRank



# Test 2: Grafico Anti-trustRank



# Test 2: Risultati

- Dai test si deduce che Anti-trust Rank è più efficace di Trustrank in assoluto (vedi test 1)

# Test 2: Risultati

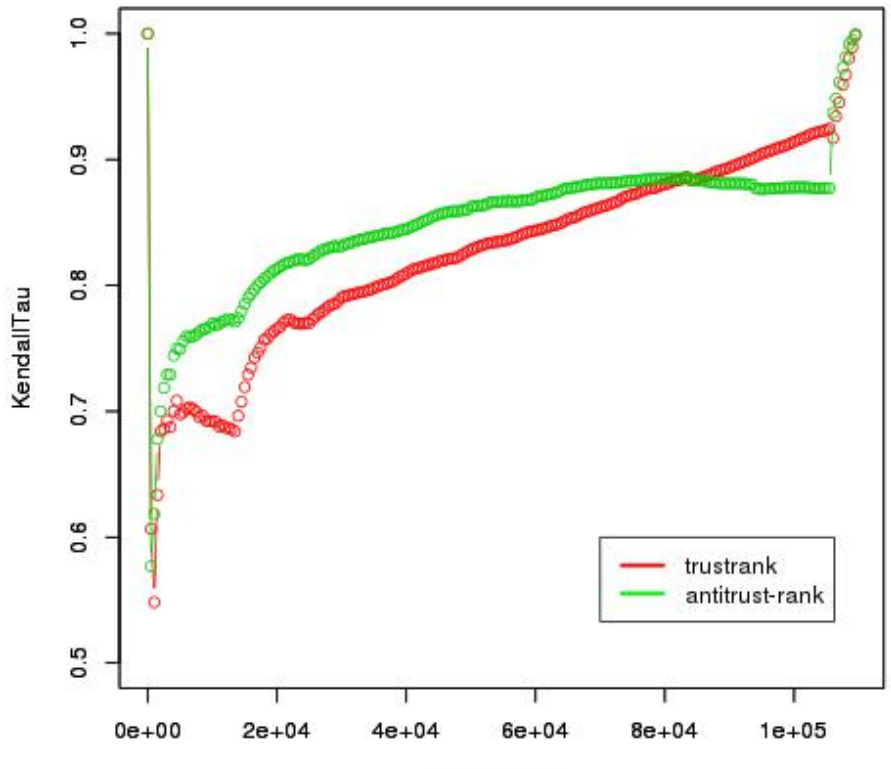
- Dai test si deduce che Anti-trust Rank è più efficace di Trustrank in assoluto (vedi test 1)
- Trustrank è più efficace ad identificare i nodi spam

# Test 3

Si è simulato il caso in cui il seedset utilizzato dagli algoritmi sia formato da nodi al limite del grafo $G$.

# Test 3

Si è simulato il caso in cui il seedset utilizzato dagli algoritmi sia formato da nodi al limite del grafo $G$.



# Test 3: Risultati

- Anti-trust Rank è meno dipendente dal vettore di preferenza che gli viene passato rispetto a Trustrank, almeno fino a quando il grafo temporaneo è al massimo l'80% del grafo completo

# Test 3: Risultati

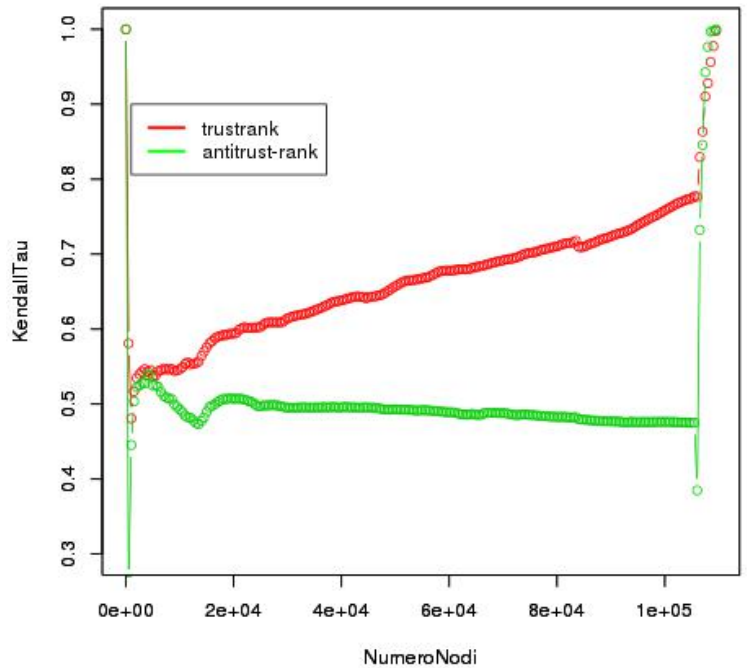
- Anti-trust Rank è meno dipendente dal vettore di preferenza che gli viene passato rispetto a Trustrank, almeno fino a quando il grafo temporaneo è al massimo l'80% del grafo completo
- Quindi anche se si usasse Anti-trust Rank in modalità online con un seedset poco pertinente questo riuscirebbe a produrre, fin dall'inizio del crawling, dei risultati molto vicini a quelli calcolati Anti-trust Rank sull'intero grafo con un altro seedset

# Test 4

Simile al test 3 ma il fattore di attenuazione $\alpha$ con cusi si calola il vettore di TrustRank sull'intero grafo e sul grafo temporaneo è impostato a 0.005.

# Test 4

Simile al test 3 ma il fattore di attenuazione $\alpha$ con cusi si calola il vettore di TrustRank sull'intero grafo e sul grafo temporaneo è impostato a 0.005.



# Test 4: Risultati

- TrustRank è meno dipendente dal fattore $\alpha$ rispetto ad Anti-trust Rank

# Test 5

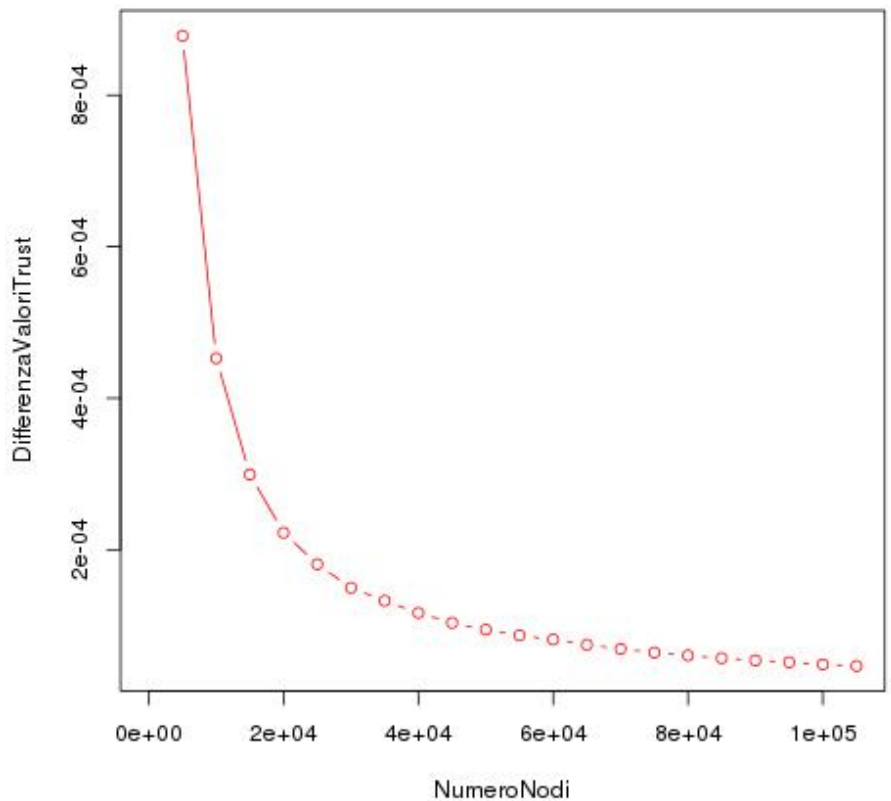
- Dal vettore di TrustRank, calcolato sul grafo temporaneo, viene calolata la differenza $\Delta_t$ tra la media $Mb_t$ dei valori di TrustRank dei nodi non spam e la media $Ms_t$ dei valori di TrustRank dei nodi spam.

$$\Delta_t = Mb_t - Ms_t \tag{3}$$

- Dal vettore di Anti-trust Rank, calcolato sul grafo temporaneo, viene calolata la differenza $\Delta_a$ tra la media $Ma_t$ dei valori di Anti-trust Rank dei nodi non spam e la media $Ms_t$ dei valori di Anti-trust Rank dei nodi spam.

$$\Delta_a = Mb_a - Ms_a \tag{4}$$

# Test 5: TrustRank



# Test 5: Risultati

- Il risultato atteso è che durante la visita il calcolo di TrustRank discrimini in modo efficace i nodi non spam dai nodi spam e quindi che $\Delta_t$ cresca durante la visita.

# Test 5: Risultati

- Il risultato atteso è che durante la visita il calcolo di TrustRank discrimini in modo efficace i nodi non spam dai nodi spam e quindi che $\Delta_t$ cresca durante la visita.
- Ma i risultati del grafico indicano che all'inizio della visita la differenza $\Delta_t$ ha un valore più alto rispetto alla differenza $\Delta_t$ calcolata sul grafo temporaneo che si ottiene ai passi successivi della visita in ampiezza.

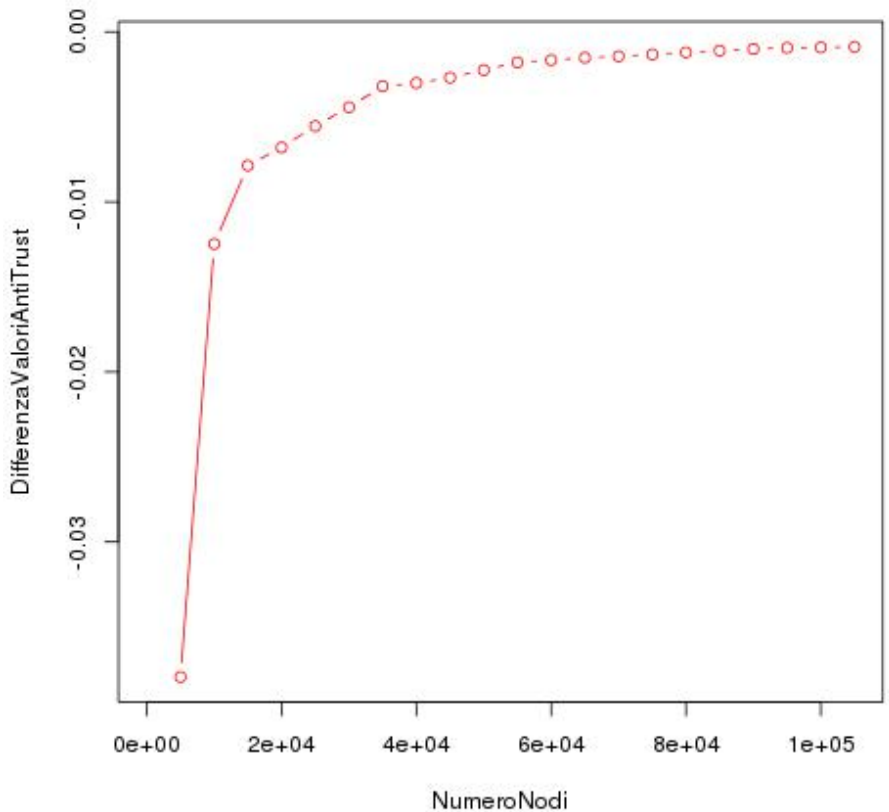
# Test 5: Risultati

- Il risultato atteso è che durante la visita il calcolo di TrustRank discrimini in modo efficace i nodi non spam dai nodi spam e quindi che $\Delta_t$ cresca durante la visita.
- Ma i risultati del grafico indicano che all'inizio della visita la differenza $\Delta_t$ ha un valore più alto rispetto alla differenza $\Delta_t$ calcolata sul grafo temporaneo che si ottiene ai passi successivi della visita in ampiezza.
- Tale comportamento implica che la distanza tra i valori di TrustRank tra i nodi non spam e spam tende ad essere meno marcata tanto più il grafo temporane $G_v$ su cui viene calcolato TrustRank cresce.

# Test 5: Risultati

- Il risultato atteso è che durante la visita il calcolo di TrustRank discrimini in modo efficace i nodi non spam dai nodi spam e quindi che $\Delta_t$ cresca durante la visita.
- Ma i risultati del grafico indicano che all'inizio della visita la differenza $\Delta_t$ ha un valore più alto rispetto alla differenza $\Delta_t$ calcolata sul grafo temporaneo che si ottiene ai passi successivi della visita in ampiezza.
- Tale comportamento implica che la distanza tra i valori di TrustRank tra i nodi non spam e spam tende ad essere meno marcata tanto più il grafo temporane $G_v$ su cui viene calcolato TrustRank cresce.
- Perciò invece di avere un andamento logaritmo i grafici sono contraddistinti da una parabola decrescente.

# Test 5: Anti-trust Rank



# Test 5: Risultati

- Anti-trust Rank calcolato verso la fine della visita dovrebbe restituire dei valori per i nodi non spam molto piccoli e per i nodi spam dei valori molto alti. Si deduce quindi che alla fine della visita , $Mb_a$ dovrebbe essere più piccola di $Ms_a$ e di conseguenza che $\Delta_a$ dovrebbe essere negativa .

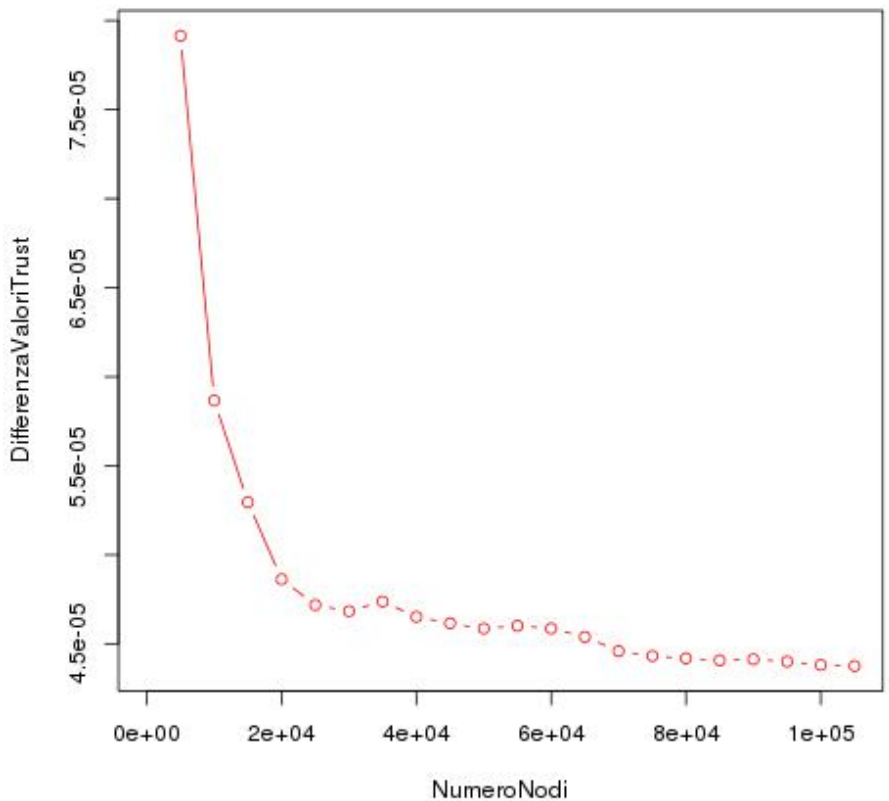
# Test 5: Risultati

- Anti-trust Rank calcolato verso la fine della visita dovrebbe restituire dei valori per i nodi non spam molto piccoli e per i nodi spam dei valori molto alti. Si deduce quindi che alla fine della visita , $Mb_a$ dovrebbe essere più piccola di $Ms_a$ e di conseguenza che $\Delta_a$ dovrebbe essere negativa .
- Anche in questo caso i risultati illustrati nei due grafici smentiscono il comportamento atteso in quanto il valore di $\Delta_a$ aumenta con l'aumentare dei nodi del grafo temporaneo su cui viene calcolato Anti-trust Rank.

# Test 6

- Dal momento che il test numero 5 ha prodotto dei risultati che indicano un comportamento diverso da quello atteso, è stato implementato questo test per verificare la correttezza dei risultati ottenuti.
- Il test è simile al test 5 ma invece di usare i valori temporanei di TrustRank e Anti-trust Rank per calolare $\Delta_t$ e $\Delta_a$, durante la visita in ampiezza, si utilizzano i valori ricavati da TrustRank e Anti-trust Rank calcolati sul grafo completo.

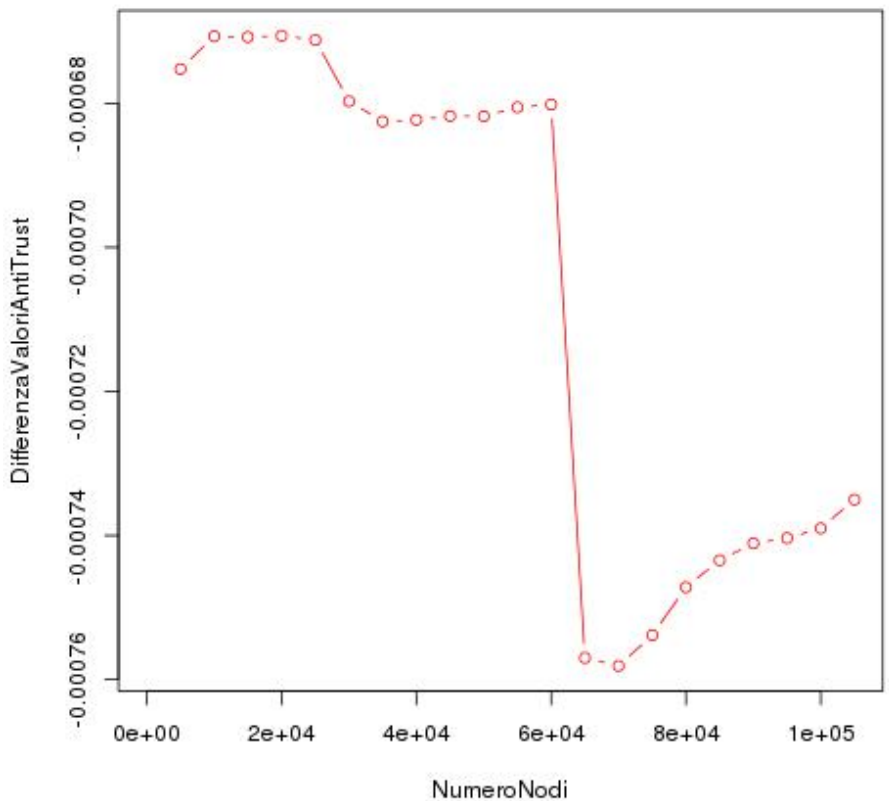
# Test 6: TrustRank



# Test 6: Risultati

- L'andamento del grafico giustifica i risultati ottenuti nel test numero 5 il quale ha un andamento molto simile.
- Il range in cui variano i valori $\Delta_t$, in questo test il range è di un ordine di grandezza più piccolo e quindi si deduce che la media dei valori di TrustRank del gruppo di nodi spam e la media TrustRank del gruppo di nodi non spam, contrariamente a quanto ci si aspetta, sono molto vicine.
- Perciò il test confermache i risultati ottenuti nel test 5 dipendono dall'algoritmo di TrustRank.

# Test 6: Anti-trust Rank



# Test 6: Risultati

- I risultati del grafico non seguano quelli del test 5
- Ma dal momento che i valori del grafico variano in un range molto piccolo possono confermare i risultati ottenuti nel test numero 5

# Conclusioni

- TrustRank e Anti-trust Rank possono essere usati in modo online in quanto approssimano abbastanza bene il loro comportamento offline
- Confrontando i due algoritmi si è evinto che Anti-trust Rank approssima, per quasi tutta la durata del crawling, meglio il comportamento offline e quindi è più indicato per essere usato durante la fase di crawling.
- Una volta identificate le pagine spam molte altre pagine potrebbero essere non scaricate.

## Sviluppi futuri

Sviluppo futuro di tale lavoro sarà la progettazione di un modulo di spam detection da inserire all'interno del web crawler BUbiNG.