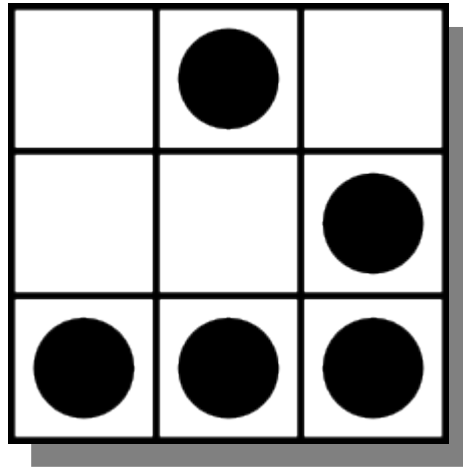


Motori di ricerca: pericolo?



Giacomo Rizzo [a.k.a. alt-os]

alt@free-os.it



Premessa

- Non è una presentazione contro Google
 - Google è solo un esempio
 - Recentemente modifica a normative privacy (da vedere)
- Si basa su dati noti:
 - Normativa privacy di Google
 - Un minimo di legge internazionale
 - Documenti accessibili tramite internet
- Obiettivo
 - Far presente alcune problematiche che fanno parte “del gioco” ma vengono spesso e volentieri dimenticate
 - Si fa riferimento ad un modello economico in cui l'informazione stessa ha un valore difficilmente quantificabile e la privacy di conseguenza

Contesto

- “3a era dell'Informazione”
 - 1a era dell'informazione: “la stampa”
 - Potere estremamente centralizzato
 - Legato ai costi d'accesso della “trasmissione”: tipografia
 - Alti costi d'accesso (costo dei libri)
 - 2a era dell'informazione: “radio e TV”
 - Potere estremamente centralizzato
 - Legato ai costi d'accesso della “trasmissione”: emittente radio-televisiva / competenze
 - Ridotti costi d'accesso (apparecchio TV)
 - 3a era dell'informazione: “internet”
 - Potere “decentralizzato”
 - Costi di accesso ridotti all'osso

Soggetto

- Il motore di ricerca
 - Svolge una funzione estremamente importante
 - Districare l'enorme quantità di informazioni reperibili in rete
 - Troppi risultati = nessun risultato
 - Accentrare l'accesso a queste risorse
 - Conoscere l'esistenza di una pagina per poterla trovare
 - “L'utente non sa quel che cerca”
 - Sono l'occhio trami il quale vediamo la Rete
 - Risultati incredibilmente diversi (fonte: Jux2)
 - A parità di query Google e Yahoo! Forniscono:
 - 3.8/10 risultati presenti in entrambi (38%)
 - 23/100 risultati presenti in entrambi (23%)
 - 4.8 di top 10 Google in top 100 Yahoo!
 - 5.4 di top 10 Yahoo! in top 100 Google

Una restrizione percettiva

- Andiamo quindi ad individuare una possibile “restrizione percettiva” (miopia)
 - Noi vediamo con “gli occhi” del motore di ricerca
 - I suoi occhi sono il suo algoritmo di ricerca
 - Se si tratta di una visione imparziale, gestita tramite meccanismi noti e pubblici il problema è contenuto
 - Rischio dell'abuso
 - E' lecito attendersi imparzialità da un'entità che ha nel profitto economico la sua ragione di esistere?
 - Se sfruttasse un modello di business basato sull'etica, sì
 - Ma in un'ottica prettamente capitalistica?
 - Abuso “ufficiale”
 - Censura tramite motore di ricerca (Cina ma non solo)
 - Opinioni altrui applicate indirettamente alla mia percezione

Una restrizione percettiva

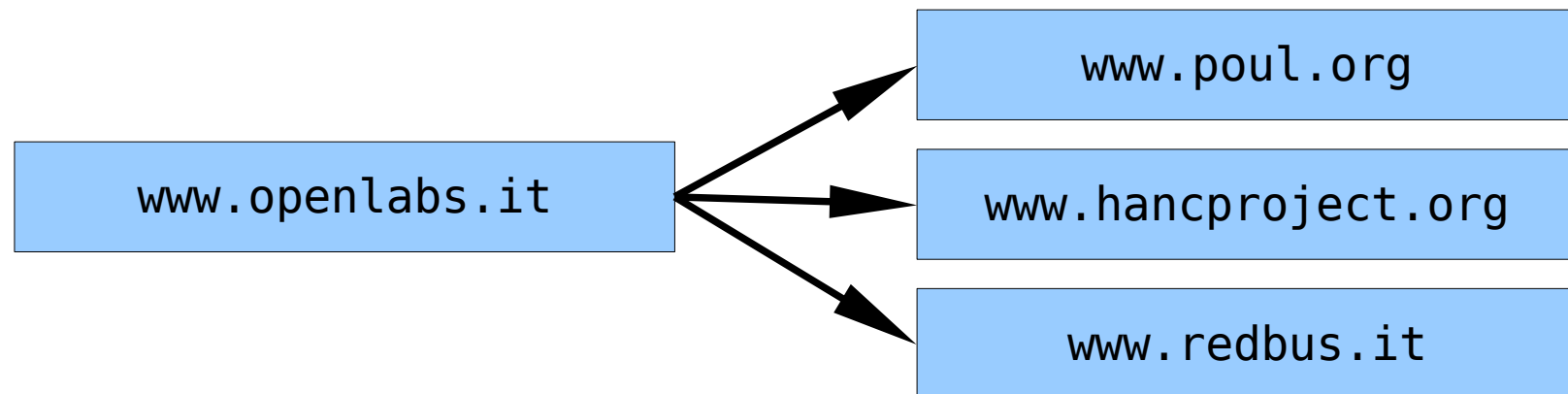
- I motori di ricerca affermano di non praticare censura (al di là di quanto previsto dalla legge) ed eventualmente di avvertire con un banner gli utenti quando questo si verifica
 - Questo accade in Cina in maniera diffusa e sistematica
 - Ma non solo. Anche in Francia e Germania Google applica una forma di censura su alcuni risultati delle ricerche
<http://www.google.de/search?hl=en&q=allinurl%3Astormfront.org>
 - Uno studio condotto in questi due paesi ha portato al risultato di 113 casi di censura di questo tipo.
 - A quando, in Italia, l'esclusione dei siti di scommesse non affiliati ad AAMS dai risultati di Google? :/
 - Tanto basta cambiare il server DNS... :/ :/

Una restrizione percettiva

- Ma la censura, è cambiata insieme alla realtà in cui viviamo, insieme al modo di usare il network
- Il meccanismo “non-democratico” di ranking di Google non porta forse ad una forma di censura oligarchica, facendo dei siti con rank più alto degli involontari censori?

Google Ranking

- Un link di A = un voto, alza il ranking, per B
- Il link di una pagina con ranking più alto, ha più valore (tentativo di dare valore all'autorevolezza di una pagina, valutata numero di link [a loro volta autorevoli o meno])
- Il problema è che presto questo meccanismo porta a generare circoli viziosi: le pagine autorevoli hanno tutte un interscambio di link che “chiude fuori” dai risultati i nuovi siti, anche oggettivamente autorevoli
- Meccanismo vulnerabile ad attacchi (molto usato dai SEO)

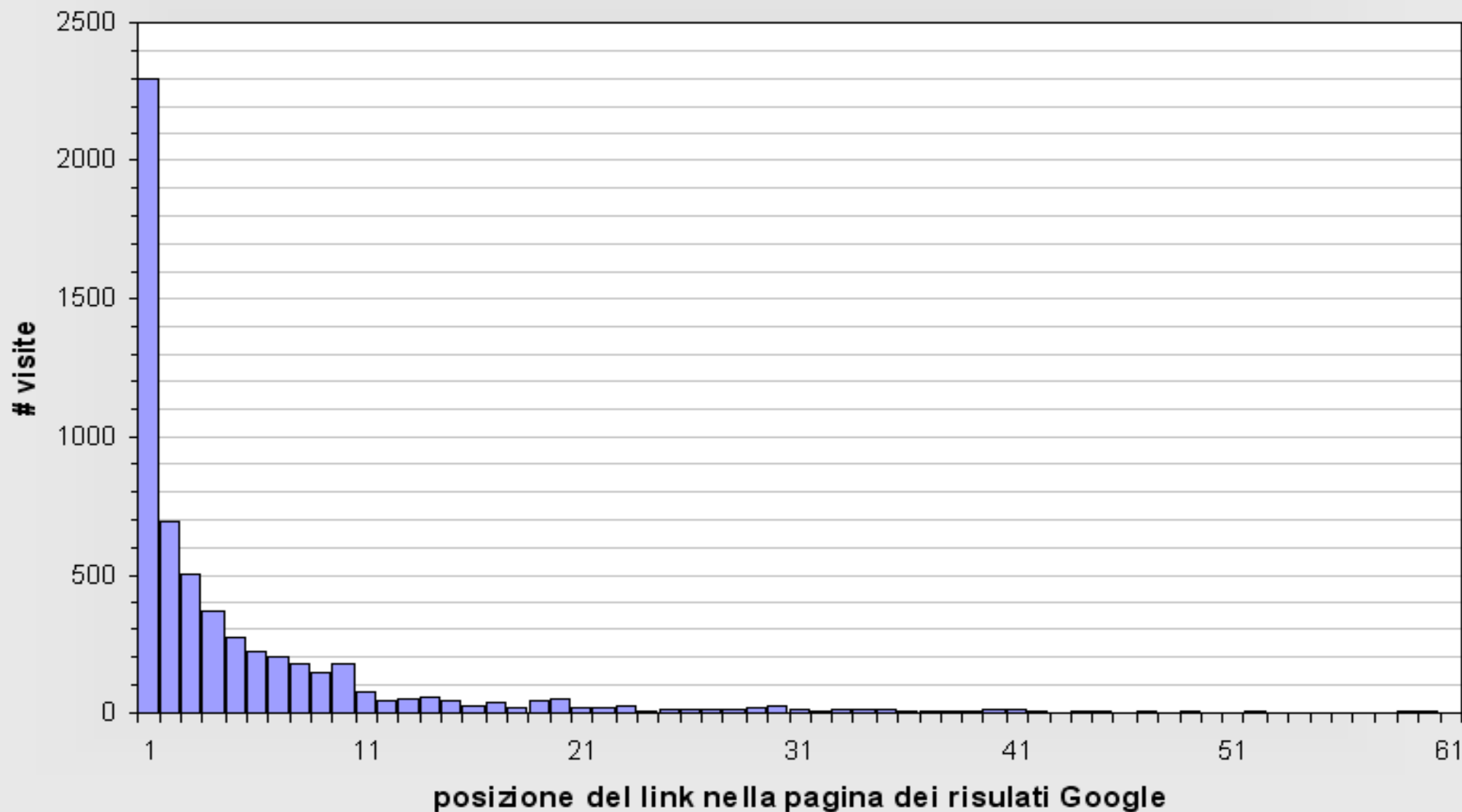


Una restrizione percettiva

- Ma la censura, è cambiata insieme alla realtà in cui viviamo, insieme al modo di usare il network
- Il meccanismo “non-democratico” di ranking di Google non porta forse ad una forma di censura oligarchica, facendo dei siti con rank più alto degli involontari censori?
 - Un sito che non raggiunge un ranking sufficientemente alto da entrare nei primi 10 posti, è come se fosse censurato, non esiste.
 - Gli utenti infatti cliccano il primo risultato. Sempre. Al più, arrivano a cliccare i primi 5-6, fino ad un massimo di 10. Gli altri, sono praticamente risultati inesistenti.
 - Uno studio di Matteo Flora e Claudio Agosti mette ben in evidenza questo aspetto...

Una restrizione percettiva

Distribuzione delle visite degli utenti in funzione della posizione del link nei risultati di una ricerca Google



Problema della “geolocalizzazione”

- Geolocalizzazione
 - Generare risultati alle ricerche in base alla localizzazione geografica della richiesta
 - Viene già implementato
 - www.google.it
 - www.google.com
 - Tecniche più avanzate
 - Sinergia con altre informazioni
 - Google Maps ed attività commerciali
 - Pubblicità
 - Riterremo più vicino il primo risultati geolocalizzato di Google Maps/Attività Commerciali rispetto al fruttivendolo sotto casa se questo non è inserito in G.Maps?

Funzionamento di un m.d.r.

- Piuttosto semplice
 - Crawling (raccolta dati)
 - Uno spider gira per la rete visitando pagine
 - Indexing (indicizzazione)
 - Tramite l'applicazione di appositi algoritmi, ad ogni pagina visitata viene assegnato un valore (detto “rank”) che ne determina la posizione in una “classifica”
 - Searching (ricerca)
 - A fronte della richiesta di un client, si effettua una ricerca all'interno della “classifica” che fornisce i risultati cercati dall'utente
- Problematiche legate al numero di accessi, all'interpretazione delle stringhe di ricerca, allo storage dei dati, alla banda utilizzata: problemi risolvibili

Funzionamento di un m.d.r.

- La differenza, oggi, non si fa più sull'algoritmo di indicizzazione e/o ranking
 - Non c'è “più nulla” da inventare (nessuna innovazione recente), gli algoritmi utilizzabili sono noti
- Ma sulla “profondità del crawling”
 - Ovvero su quanta quota parte della Rete si riesce ad indicizzare
 - Legato a banda, storage => investimenti economici
 - Legato al tempo a disposizione (la Rete è immensa)
 - Problema del “sommerso” (pagine che non presentano link da altre pagine. Raggiungibili solo su segnalazione da parte degli utenti (che le scrivono))

Motori di ricerca a confronto

- Ask.com
 - 2,5 miliardi di pagine indicizzate
 - Profondità: > 101Kb/pagina
 - 13 milioni di ricerche al giorno
 - Motore molto giovane, buone intenzioni, buona tecnologia
 - Rimane da capire quanto in profondità riusciranno ad andare con il crawling
- Microsoft Live-Search
 - 5.0 miliardi di pagine indicizzate
 - Profondità: 150Kb/pagina
 - 28 milioni di ricerche al giorno
 - Sviluppato un nuovo motore perché si è trovata incastrata tra Inktomi e Overture, entrambi acquistati da Yahoo!

Motori di ricerca a confronto

- Yahoo!
 - 4.2 miliardi di pagine indicizzate (stima)
 - Profondità: 500Kb/pagina
 - 60 milioni di ricerche al giorno
 - Ha compiuto pesanti investimenti per ampliare le potenzialità della propria piattaforma, acquistando altri motori quali
 - Altavista
 - Overture
 - Alltheweb
 - Inktomi
 - Hanno espresso chiaramente, più di una volta la volontà/necessità di produrre utili tramite la vendita di **ranking** (da verificare)

Motori di ricerca a confronto

- Google
 - 8.1 miliardi di pagine indicizzate
 - Profondità: 101Kb/pagina
 - 91 milioni di ricerche al giorno
 - La principale fonte di referrer per quasi tutti i siti web (75%)
 - Dichiarata come “missione”:
 - La nostra missione è quella di raccogliere e organizzare le informazioni di tutto il mondo, e stiamo dedicando le nostre risorse a questo scopo.

<http://www.google.it/support/bin/answer.py?answer=454&topic=369>

- “E salvare il mondo dalle orde assassine di Alfa Centauri. E poi via, verso nuovi mondi e nuove avventure...”
- “E poi c'era la marmotta che confezionava la cioccolata”

Motori di ricerca a confronto

- Esistono altre tipologie di motori di ricerca: i così detti meta-motori
 - Sfruttano i database (e quindi il crawling) degli altri motori di ricerca, combinandoli per fornire un risultato più ampio
 - Comportamento proibito da Google
 - Juk2
 - Ricerche comparative con raccolta dati statistica
 - IxQuick (Olanda)
 - Ricerche comparative e niente logging (dal 2006)

La piattaforma di Google

La piattaforma che Google mette a disposizione “gratuitamente” ai propri utenti è davvero immensa:

- Blogger/Blogspot.com
- Search
 - ricerca personalizzata
 - homepage personalizzata
 - my search history
- Google Groups
- Google Toolbar
- Web Search Features
 - Images
 - Movies
 - Books
 - Stocks
 - PhoneBook
 - Calculator
 - Language tools (traduzioni)
 - Convertitore di valuta/unità di misura
 - Definizioni

La piattaforma di Google

- Google Web Accelerator
- Google Alerts
- Froogle
- Google Video
- YouTube
- Google Docs&Spreadsheets
- Google News
- Google Scholar
- Google Talk
- Google Desktop
- Google Directory
- Google Code
- Google AdWords/AdSense
- Google Pages
- Google Pack
- Google Sketch Up
- Picasa / Picasa Web
- Google Earth/Maps
- Google Moon
- Google Mars
- Google Local
- Google Checkout

La piattaforma di Google

- Google Mail
- Google Catalogs
- Google Sets
- Google Store
- Google Base
- Google Finance
- Google Reader
- Google Trends
- Google Mobile/SMS
- Google Calendar
- Google Patent Search
- Google Books
- Google Blogs
- Orkut
- In arrivo
 - Google Transit
 - Google Music Trends
 - Google Ride Finder
 - Google Notebook
 - Google Gulp (hehehe)
 - Google Phone

I dati raccolti (comuni)

- Una piattaforma così vasta, significa una potenziale raccolta di informazioni di tipologie estremamente diverse, anche in maniera molto approfondita.
- Cosa potrebbe comporre l'elenco dei dati in mano a Google, nel caso in cui li usasse per fare profilazione?
 - Una parte di dati, il nostro browser li invia comunque, o sono facilmente ricavabili:
 - Indirizzo IP
 - Localizzazione Geografica , Tipologia di connessione
 - Lingua del browser
 - Sito di provenienza (referrer)
 - Piattaforma operativa (sistema operativo, browser)
 - Data e ora delle connessioni
 - Cookie di identificazione *

I dati raccolti (comunicati)

- Ma sono i dati forniti “volontariamente” dagli utenti ad essere quelli più interessanti.
- A seconda dei servizi di Google che utilizziamo potrebbero ricavare:
 - Informazioni finanziarie
 - Homepage personalizzata (quotazioni di borsa), Google Finance, Google Check Out
 - Elenco dei nostri “contatti”, più o meno vicini
 - Google Mail, Google Talk, Google Groups, Google Mobile
 - Nostri messaggi (posta, IM, blog, ...)
 - Google Mail, Google Talk, Orkut, Google Groups, Blogger.com
 - Interessi personali
 - Google News, Homepage personalizzata (feeds RSS), Google Reader, Google Mail, Google Search

I dati raccolti (comunicati)

- Siti visitati
 - Google Ads *, Homepage personalizzata (feeds RSS), Google Toolbar, Google Web Accelerator, Google Desktop
- Informazioni anagrafiche personali
 - Ogni account (serve fornirli falsi?), Google CheckOut
- Frequenza e tipologia di utilizzo del computer
 - Google Desktop, Google Pack, Google Talk, Google Toolbar, Google Web Accelerator
- Files sul sistema installato
 - Google Desktop, Google Pack
- Attività
 - Google Calendar, Google Mail, Google Mobile, Google Maps/Earth

I dati raccolti (ricavati)

- Una serie di dati possono poi essere ricavati, incrociando dati delle altre tipologie
 - Attività sul Web
 - Google Adsense, Google Toolbar, Google Web Accelerator
 - Click sui risultati delle ricerche effettuate *
 - Informazioni geografiche
 - Geolocalizzazione IP
 - Google Maps/Earth
- In alcuni casi, Google consente, durante l'attivazione dell'account, di dichiarare esplicitamente la non disponibilità a far incrociare i propri dati.
- Non sempre è prevista questa scelta.

Qualche curiosità... cookie

- Il Cookie di Google
 - Cos'è un cookie?
 - Un cookie (biscottino) è un file che il webserver chiede al browser di salvare in locale, contenente alcuni dati testuali.
 - Viene normalmente utilizzato per gestire le sessioni, o per consentire ad un utente di trovarsi già loggato alla successiva visita ad un sito web
 - Quanto dura un cookie?
 - La dove non specificato diversamente dal webserver, viene eliminato quando si chiude il browser
 - Essendo la data rappresentata con 32 bit, può arrivare al massimo al 18 gennaio 2038
 - Quando scade “il cookie di identificazione” di Google?
 - Il mio il 17 gennaio 2038, alle 20:13:40. Il vostro?

Qualche curiosità... cookie

- Il Cookie di Google
 - In più, c'è il cookie di Google Analytics (statistiche d'accesso) che scade nel 2038, ed è presente su un numero impressionante di siti (non collegati a Google direttamente)
 - Non che gli altri motori di ricerca non facciano altrettanto:
 - Yahoo!: 2037
 - Live.com/msn.com: 2021
 - Ask.com: 2038
- Precisazione: In se, la presenza di questo cookie non significa nulla di male, ma può costituire il collante per mettere in relazione tutte le informazioni ricevute tramite le diverse fonti (il cookie viene inviato ogni qual volta ci si connette ad uno di quei server)

Qualche curiosità... tracking

- Tracking dei risultati
 - Per un motore di ricerca è molto interessante sapere quali risultati di una ricerca gli utenti cliccano
 - Questo infatti può essere usato per migliorare le risposte alle ricerche (se tutti cliccano il secondo risultato...)
 - Si pone però un problema:
 - Quando clicco un link (ed i risultati di una ricerca sono dei links), il link cliccato non viene comunicato al server dove si trova, perché il nostro browser contatta direttamente la destinazione per ottenere la pagina da visualizzare.
 - Come si può fare allora? Si dovrebbe trovare il modo di far “ripassare” il client per le pagine del motore di ricerca prima che si diriga verso la reale destinazione...
 - Ogni motore di ricerca ha trovato una soluzione diversa...

Qualche curiosità... tracking

- Yahoo!
 - Fino a qualche tempo fa, usava un banale redirect
 - Il link puntava ad una pagina di Yahoo! che poi reindirizzava il browser verso la destinazione
 - Oggi?
- Google
 - Introduce un semplice javascript che, al click su un link, richiede un'immagine inesistente a Google, passandogli così il numero del risultato scelto ed il cookie
 - Perché solo Internet Explorer? boh...
- Ask e Live.com
 - Usano un sistema molto simile a quello usato da Google, ma funziona anche con gli altri browser.

Usi dei dati raccolti

- Ma che se ne fanno dei dati così raccolti? La normativa privacy di Google dice che li usano per:
 - Migliorare il servizio
 - Fatturazione a terzi (Google Ads)
 - Fornire contenuti personalizzati (pubblicità, risultati)
 - Fornitura di statistiche “aggregate e non personali”
- Poi ci si protegge dagli abusi (logico)...
 - Protezione del network e dei servizi
 - Protezione da frodi e/o danni imminenti
- Infine ci sono i procedimenti legali da parte delle autorità inquirenti.
 - Qui siamo al sicuro, no?

Usi dei dati raccolti

- No.
- Secondo i canoni del diritto internazionale, la legge di tutela della privacy che si applica ai dati raccolti è quella del paese in cui risiede il responsabile del trattamento.
- Nel caso di Google Italia, Mountain View, California, Stati Uniti d'America.
- Questo fa scattare un po di preoccupazione. Nomi come Patriot Act e “Guerra al terrore” suonano piuttosto familiari alle nostre orecchie.
- Ma non saltiamo sulla sedia solo perché non conosciamo adeguatamente la normativa statunitense.
 - Le NSL

Le NSL

- NSL = National Security Letter
- Introdotte nel 1970 per combattere terrorismo e spionaggio internazionale
- Sono state estese nel 2001 dal Patriot Act, dando pieni poteri all'FBI, che le può emettere a propria discrezione, senza necessità di mandato giudiziario, anche nei confronti di persone che non hanno commesso alcun reato.
- Consentono di raccogliere, per quel che riguarda la Rete, tabulati telefonici, log di server (email, web, ...) e via dicendo. Non però il contenuto dei messaggi.
- Si tratta di documenti sottoposti a segreto. Chi li riceve non può parlarne, pena, il carcere.

Le NSL

- In 4 anni (fino al 2006) ne sono state emesse 30.000, ognuna delle quali può riguardare più di una persona (i quali, ovviamente, non lo sapranno mai)
- Nonostante il Patriot Act preveda una scadenza per altre tipologie di misure, le NSL non rientrano tra di esse.
- L'esistenza dei dati raccolti, non è neppure limitata alla durata dell'indagine: infatti Ashcroft ha modificato una delle linee guida dell'FBI che prevedeva la distruzione di tutti i dati raccolti nel momento in cui questi non fossero più rilevanti ai fini delle indagini o il soggetto profilato fosse in qualche modo scoperto innocente.
- Oggi questi dati vengono immagazzinati e messi in relazione in prospettiva di nuove possibili indagini, anche non interne all'FBI stessa.

Sicurezza o privacy

- Di fronte a violazioni del diritto alla privacy di questa entità, l'abuso è parte stessa della violazione
 - Sapendo che potrebbero raccogliere anche i vostri dati, andreste tranquillamente a visitare siti “scomodi” (sito web di Al-Jazeera)?
 - Il timore stesso porta ad una forma di censura, quindi ad un abuso
- Dopo l'attentato dell'11 settembre, le misure di sicurezza di questo genere si sono moltiplicate
 - Non solo negli USA
 - Si tratta di misure inutili
 - Possono addirittura diventare controproducenti
 - Data retention / dataveillance / telesorveglianza

Dataveillance

- Alcuni problemi intrinseci
 - Alza i costi di chi deve effettuarla
 - Riduce di conseguenza la qualità del servizio (possibile)
- Dall'11 settembre 2001, molte restrizioni alla privacy degli utenti sono state giustificate dalla “guerra al terrorismo”
- La carta dei diritti fondamentali dell'Unione Europea, sancisce che la protezione dei dati personali è un diritto della persona: “noi siamo le nostre informazioni”
- L'attuale tendenza è quella di ridurre ulteriormente la quantità di libertà concessa ad ognuno, seguendo il motto:
 - “Tanto non hai nulla da nascondere”

Dataveillance

- Si può arrivare molto rapidamente a ledere la libertà
 - di comunicazione
 - di espressione
 - di circolazione
 - sul diritto alla salute
 - sulla condizione di lavoratore
 - sull'accesso al credito e alle assicurazioni
 - ...
- Questa approfondita conoscenza, non è limitata agli enti pubblici che richiedono il controllo e l'intercettazione, ma anche a tutti coloro che la effettuano!
 - Recente caso dei call-center che chiamano gli utenti

Dataveillance

- Ma la cosa peggiore è che è l'intercettazione generalizzata è di fatto inutile:
 - Non è in grado di sventare attentati (bisogna sapere cosa cercare)
 - Al limite è in grado di punire a posteriori, ma si tratta di una tattica perdente in partenza
 - Le misure sono sproporzionate rispetto alla reale minaccia ed ai risultati ottenibili
 - Non c'è una adeguata regolamentazione delle modalità di ritenzione dei dati sul territorio nazionale
 - Non ci sono regolamentazioni sull'uso dei nostri dati in altri stati.
 - Accordi internazionali e l'ONU dovrebbero garantire un rispetto condiviso

Dataveillance

- Ma la cosa peggiore è che è l'intercettazione generalizzata è di fatto inutile:
 - Eventuali malintenzionati riescono comunque a non essere intercettati
 - Mezzi tecnologici adeguati
 - Protocolli non intercettabili (p2p, Tor...)
 - Crittografia (sssth!!)
 - Proxy anonimi
 - Luoghi adeguati
 - Internet Cafè
 - Access Point non cifrati
 - Sistemi compromessi
 - Le intercettazioni generalizzate richieste dal decreto Urbani /Pisanu colpiscono i cittadini ma non i “cattivi”.
Che senso ha?

Dataveillance

- Può addirittura rivelarsi un boomerang!
 - Cosa succede se i dati raccolti finiscono nelle mani sbagliate?
 - Social Engineering
 - Entità di raccolta non fidate (Internet Cafè??)
 - Compromissioni di sistemi
 - Esempio banale: la riservatezza dei dati dei passeggeri di un aereo, è essenziale per evitare che un determinato volo venga scelto per un attentato in base alla presenza a bordo di una determinata persona o di un determinato gruppo religioso, magari identificabile in base alle abitudini alimentari rilevate dalla richiesta di un pasto.
 - Non è vero che non succede: nel 2005, sono stati "trafugate" le informazioni relative a 52.000.000 di utenti MasterCard

Dataveillance

- L'articolo 15 della Costituzione Italiana garantisce la libertà e la segretezza delle comunicazioni.
 - Art. 15: La libertà e la segretezza della corrispondenza e di ogni altra forma di comunicazione sono inviolabili. La loro limitazione può avvenire soltanto per atto motivato dell'autorità giudiziaria con le garanzie stabilite dalla legge.
- “L'intera area della comunicazione elettronica però viene lasciata sguarnita da garanzie costituzionali adeguate.” (ex. Garante per la Privacy, S. Rodotà)
- La Guerra al Terrore viene condotta contro un nemico invisibile. Dovremo avvallare qualunque tipo di controllo di massa? E fino a quando durerà tutto questo? La Guerra al Terrorismo non ha una data di inizio ufficiale, e non ne avrà una di termine.

Altri problemi

- La pubblicità mirata
 - Siamo preparati a resistere ad una pubblicità che ci parli direttamente, conoscendo i nostri gusti e le nostre inclinazioni caratteriali?
 - Ed i minori (tv)?
- La profilazione
 - Può avvalersi di importanti contributi esterni
 - Aziende come ChoicePoint possono portare a mettere in relazione una quantità di dati impressionante:
 - test antidroga
 - Precedenti insolvenze bancarie
 - informazioni mediche
 - ...
 - In mano a chi sono questi dati? Come vengono raccolti?

Conclusione

“Il prezzo della libertà è una costante, continua vigilanza, e chi ceda parte della propria libertà in nome di una maggiore sicurezza non merita ne l’una, ne l’altra...”

(Benjamin Franklyn)