Privacy e anonimato nel web odierno

“Anonimato” e “Privacy”, due termini abusati e piuttosto fumosi.

Ma dietro questi termini, e a tutta la fuffa che si trova in giro, si celano alcuni assiomi della rete come la conoscevamo; questi assiomi sono ancora validi?

Il “nemico” e' il controllo. Dobbiamo comunque distinguere ed avere sempre presente che esistono due tipi di controllo RADICALMENTE diversi del nostro uso della rete.

Uno da parte delle autorita' statali, il cui scopo e' identificare LE AZIONI della singola PERSONA FISICA.

Uno da parte delle corporation che costruiscono il web commerciale – tutte assimilabili ad aziende di pubblicita' – che sono interessate al PROFILO psico-comportamentale dell'UTENTE. L'identificazione della persona fisica e' assolutamente secondaria

I secondi vogliono vendervi dei prodotti

I primi vogliono controllare le vostre azioni

Entrambe forme di controllo e di limitazione della liberta', ma in modo profondamente diverso.

Le informazioni raccolte nel costruire il profilo commerciale di un utente possono spesso -non sempre!- essere di grandissimo aiuto nell'identificarlo.

(E, a parte casi rari, le corporation COLLABORANO con le autorita').

Come si attua il controllo degli utenti sul web? Che livello di anonimato e di privacy ha davvero un utente?

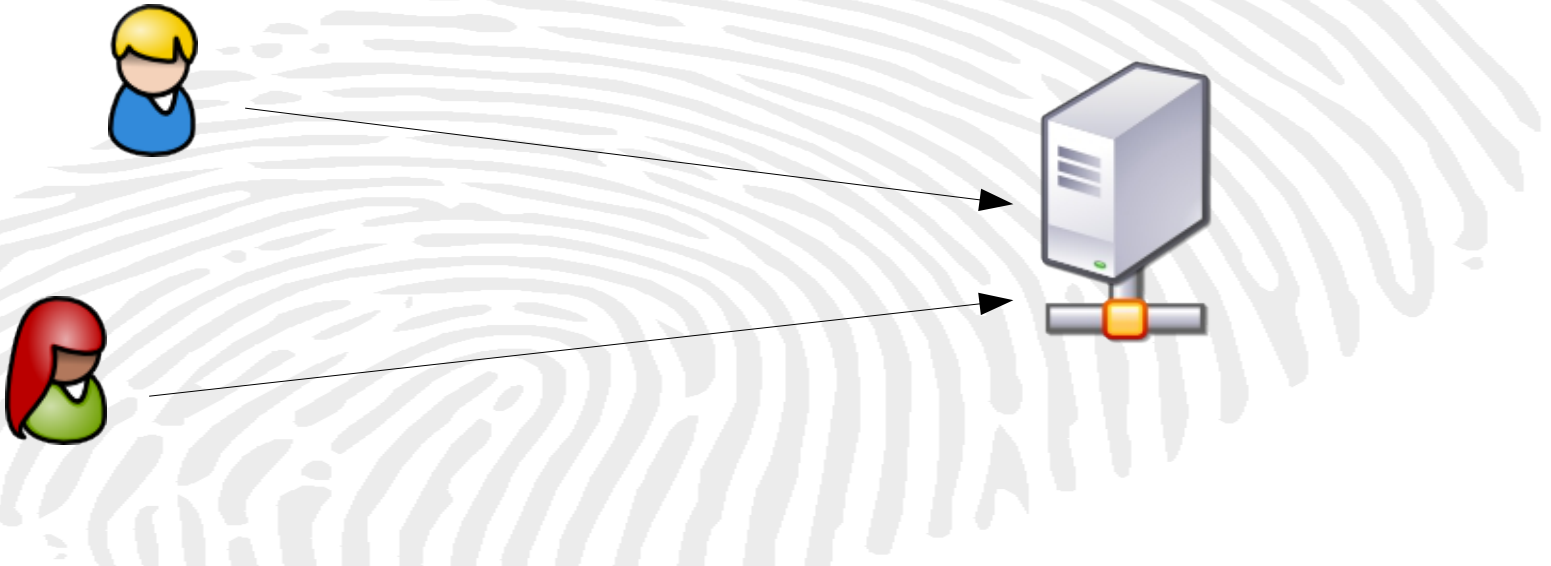
Nel prosieguo di questo seminario passero' in rassegna alcune delle principali tecniche di identificazione dell'UTENTE, utilizzate nel mondo reale da quasi qualsiasi sito web commerciale.

Con una premessa:

il concetto di privacy che ho io non è
lo stesso che ha mio padre ed è diverso anche
da quello di un ragazzo di quattordici anni.
Sei anni fa nessuno voleva che le proprie informazioni
personali fossero sul web, oggi il numero delle persone
che rende disponibile il proprio cellulare
su Facebook è impressionante.

Part II: Techniques

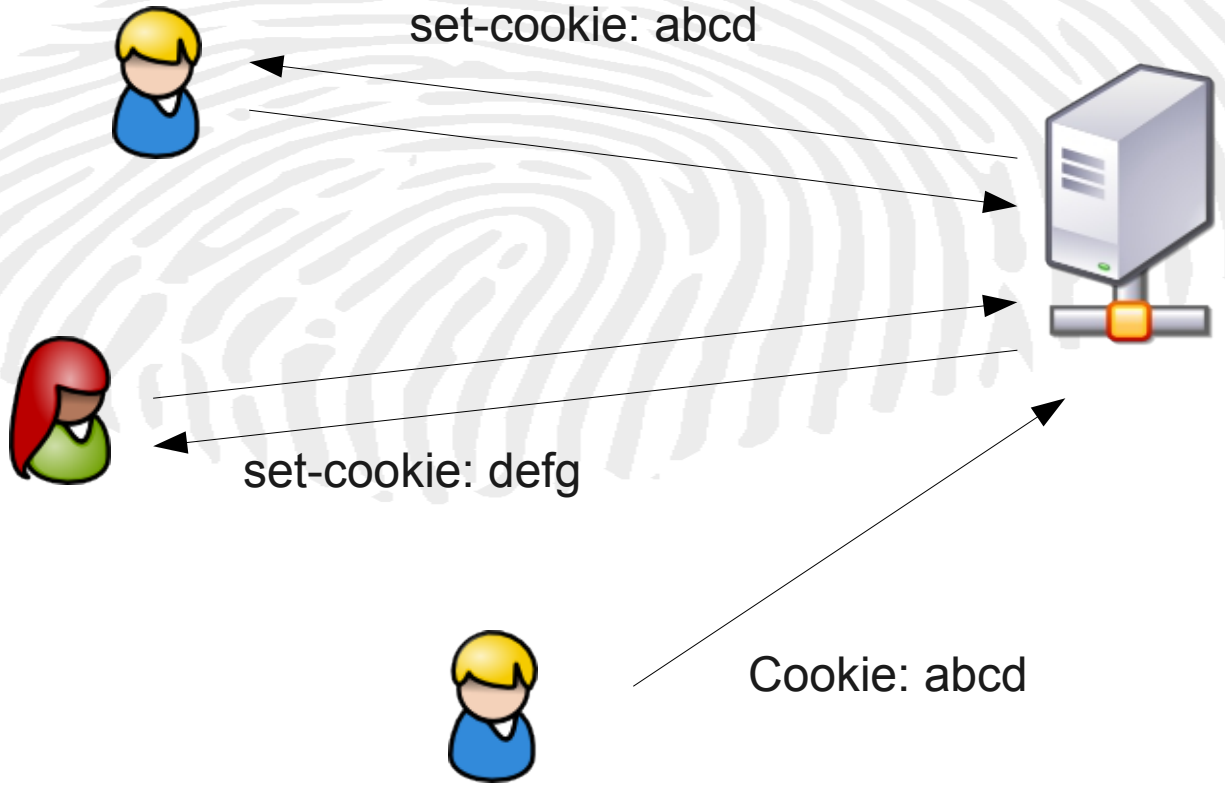
HTTP e' stateless, quindi due richieste successive non sono correlate



Come identificare le richieste successive dei singoli utenti?

La risposta (Netscape, 1994, RFC 2109, 1997): cookies

Estensione ad HTTP che permette di conservare lo stato



Per difendere gli utenti: same origin policy (SPO).

Un cookie per .foo.org puo' essere inviato per richieste provenienti da *.foo.org ma NON a richieste che provengono da bar.org

Questo crea una sandbox che permette di confinare l'identificazione dell'utente al singolo sito

Entra in gioco HTML:

```
<script src="http://www.bar.org/trackme.js" />
```

Questa richiesta, contenuta in una pagina servita da foo.org, usa le regole SOP di bar.org

bar.org e' in grado di tracciare la navigazione su foo.org:

```
… (raccolta informazioni)
var trackingRequest =
   'http://bar.org/trackimg.jpg?id=' +cookie
   + '&screenres=' + res + '&ref=' + referer + ...;
var img = new Image();
img.src = trackingRequest;
```

Se un certo numero di siti includono questo script, bar.org e' in grado di tracciare in modo completo il vostro utilizzo della rete. Questo viola effettivamente la vostra privacy dell'utente, in quanto bar.org e' in grado di raccogliere informazioni sulla persona e i suoi gusti.

Esistono script di tracciamento onnipresenti:

QUALSIASI strumento di Analytics
QUALSIASI network pubblicitario
{Facebook, Yahoo, ...} APIs

vedremo dopo che qualsiasi inclusione onnipresente e' potenzialmente dannosa.

Disabilitare i cookie, almeno quelli di terze parti. Serve?

Da' un bel senso di sicurezza, ma e' anche inutile.

La gestione dei cookie e' facile, gli utenti possono cancellarli facilmente e il problema di privacy che pongono e' noto.

Sono quindi lo strumento di base per identificare univocamente un utente, ma di per se' non bastano a garantire una identificazione pervasiva e totale.

“Difetti” dei cookie:

Solo 4 kb

Noti al grande pubblico

Facili da gestire ed esistono molti sistemi di controllo user-side

“Supercookies”

Vengono cosi' chiamati una serie di meccanismi di conservazione dello stato su HTTP alternativi ai cookie tradizionali.

In generale sono piu' capienti e flessibili dei cookies, e soprattutto meno noti e gestibili dall'utente

Alcuni browser permettono interazioni di basso livello se si usano controlli aggiuntivi (XPCOM, ActiveX).

```
//poc using io.js JS XPCOM interface
//http://kb.mozillazine.org/Io.js
var A = new FileIO();
var filename= 'path/to/my/supercookie';
var sc=A.read(filename, 'UTF-8');
if(sc)
   alert('I have a supercookie: '+ sc);
else{
   sc=createSuperCookie();
   if(A.write(filename, sc ))
      alert('New supercookie written!');
}

var I = new Image;
I.src='http://foo.org/trackme.jpg?sc=' + sc + '&loc='
   +  document.location;
```

Questi supercookies browser-specifici sono molto potenti e facilmente dissimulabili.

Pero' significa che serve un'implementazione specifica per ogni browser e non tutti i browser li supportano.

Serve un componente aggiuntivo cross-browser, cross-platform, e possibilmente con una singola implementazione.

Flash Player corrisponde all'identikit, e ha un suo sistema di supercookies:

I famosi Flash Local Shared Objects

Perche' sono considerati il Sacro Graal del tracking online?

Possono contenere fino a 100 KB di dati senza chiedere conferma

Attivi per default e sconosciuti al grande pubblico

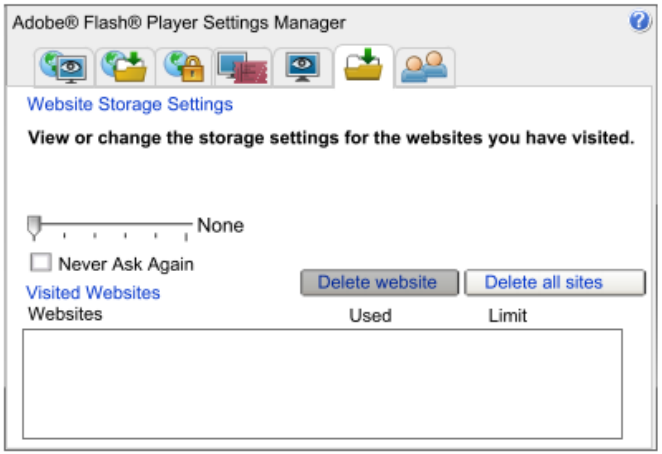
Difficilissimi da gestire, e totalmente indipendenti dai cookies.

Implementano la stessa SOP dei cookie tradizionali, con qualche sottile differenza.

http://www.macromedia.com/support/documentation/en/flashplayer/help/settings_manager07.html

Flash Player Help

Website Storage Settings panel



Note: The Settings Manager that you see above is not an image; it is the actual Settings Manager. Click the tabs to see different panels, and click the options in the panels to change your Adobe Flash Player settings.

The list of websites above is stored on your computer only, so that you can view or change your local storage settings. Adobe has no access to this list, or to any of the information that the websites may have stored on your computer.

Use this panel to specify storage settings for any or all of the websites that you have visited. The list of Visited Websites displays the following information for each website:

- The name of the website
- The amount of disk space the website has used to store information on your computer
- The maximum amount of disk space the website can use before requesting additional space

Ogni browser sul vostro sistema condivide gli LSO di flash e le impostazioni di privacy.

Uso tipico: viene usato un micro-script flash per 'reinserire' un cookie che e' stato rimosso

```
local_data = SharedObject.getLocal("cookie");
if(!local_data.cookie){
  local_data.data.cookie =
     StringUtils.generateRandomString(10);
  local_data.flush();
}
//set the http cookie to “abcde”
//Using ExternalInterface + JavaScript
```

Le “Flash alternatives” hanno ovviamente il proprio sistema di storage permanente sul client.

Es. i supercookie di Silverlight (Netflix li usa per identificare i dispositivi per il suo DRM!!!)

HTML 5, il “web del futuro”

Giustamente prevede un meccanismo avanzato di salvataggio di informazioni lato client, chiamato DOMStorage.

Utilissimo per creare applicazioni web ricche e scalabili, e facilmente funzionanti offline.

Riduce l'uso di banda per applicazioni mobile.

Ma, dal draft del W3C sullo Web Storage:

“A third-party advertiser (or any entity capable of getting content distributed to multiple sites) could use a unique identifier stored in its local storage area to track a user across multiple sessions, building a profile of the user's interests to allow for highly targeted advertising.
In conjunction with a site that is aware of the user's real identity (for example an e-commerce site that requires authenticated credentials), this could allow oppressive groups to target individuals with greater accuracy than in a world with purely anonymous Web usage”

DOMStorage: sessionStorage – session specific globalStorage[domain] – persistent, SOP

```
sessionStorage['id']=..//some random string

//on subsequent request
alert(sessionStorage['id]);
```

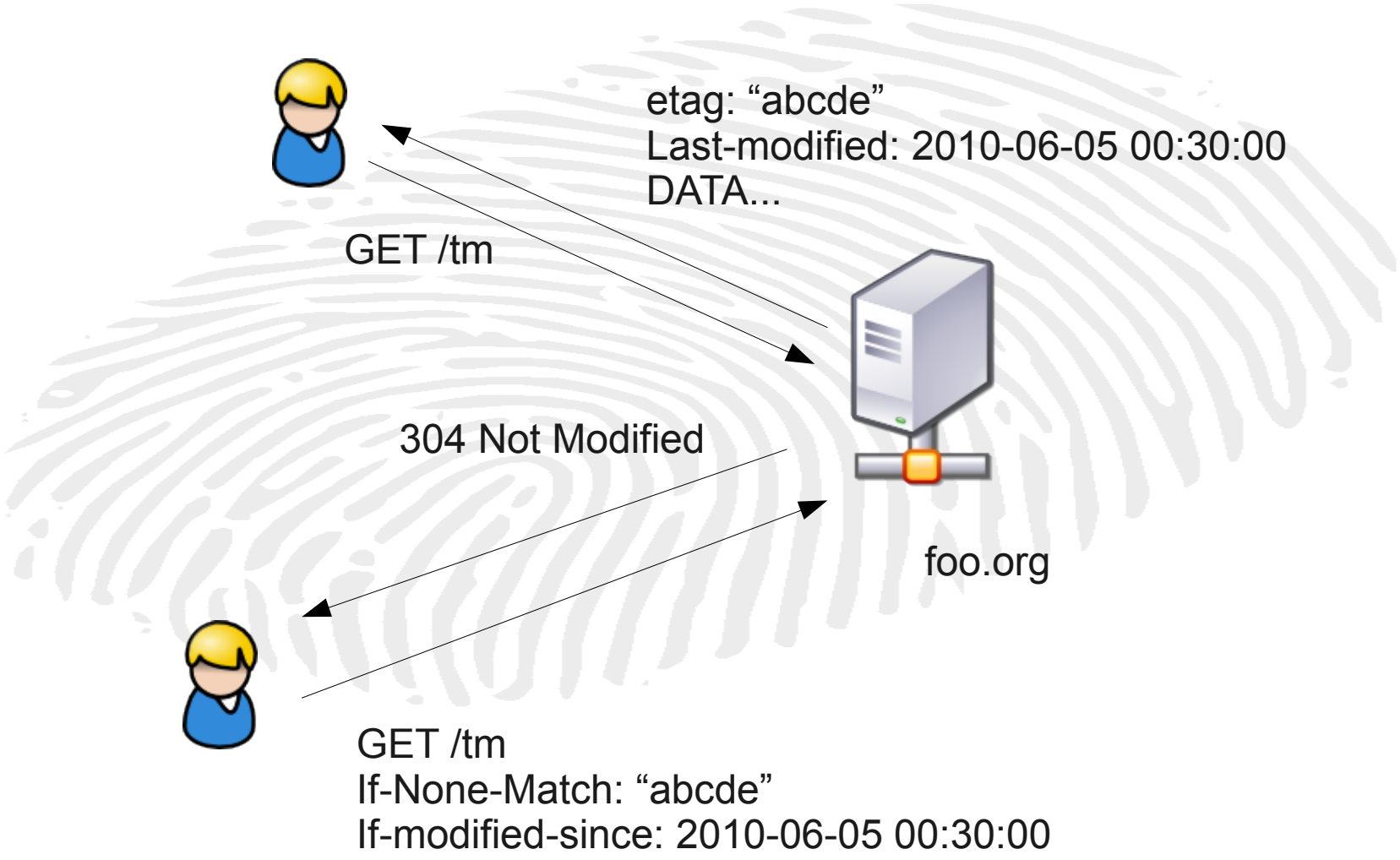
DOM Storage e' profondamente diverso dagli altri meccanismi di cookie e potenzialmente permette usi molto piu' fantasiosi.

Web cache

La cache web puo' essere usata come un sostituto dei cookie.

Ogni browser supporta il meccanismo di cache HTTP costituito da Etag/Last-Modified

Si puo' rovesciare la logica dell'etag (indentificativo univoco di uno stato di una risorsa) e usarlo per identificare l'utente in piu' visite successive (identificativo univoco dell'utente)


etag: “abcde”
Last-modified: 2010-06-05 00:30:00
DATA...
GET /tm
304 Not Modified
foo.org
GET /tm
If-None-Match: “abcde”
If-modified-since: 2010-06-05 00:30:00

L'angolo della paranoia:
I CDN sono l'ultima moda del web.

Tutti i tool per misurare le performance dei siti consigliano l'uso dei CDN, e Google dice che iniziera' a penalizzare i siti “lenti”.

Sempre piu' siti useranno i cdn per servire le librerie js.

2+2...

Altra via: cache di un contenuto univoco, da leggere poi con uno script js, via XHR o in vari altri modi.

Se proprio fallisce qualsiasi metodo di identificazione univoca (come abbiamo visto, basta che sia possibile registrare qualsivoglia informazione sul client)...

Fingerprinting!

L'idea e' di recuperare abbastanza informazioni su un utente da renderlo unico, e quindi identificarlo.

In sostanza, e' davvero lo stesso procedimento usato nell'analisi delle impronte digitali.

Lasciatevi annoiare (altri) 5 secondi.

Quanti bit di informazione servono per identificare un utente?

$$n = \log_2 N$$

Bit contenuti in una singola informazione

$$b = \log_2 P^{-1}(x)$$

Esempi: Il genere da' 1 bit; Residente a roma: 21 bit

Come si raccolgono le informazioni?

Dai dati che il nostro sistema fornisce ai server.

Primo esempio: browser history (ancora una volta, dati presenti sul vostro computer)

```
var iframe=document.createElement(”iframe”);
iframe.style.visibility = 'hidden';
document.body.appendChild( iframe );
iframe.doc=iframe.contentDocument;//Not in IE!
iframe.doc.open();
iframe.doc.write('<style>');
iframe.doc.write("a{color: #000000; display:none;}");
iframe.doc.write("a:visited {color: #FF0000; display:inline;}");
iframe.doc.write('</style>');
iframe.doc.close();
var uri= …// array of urls
for( var i in uri){
   var a=iframe.doc.createElement("a");
   a.id= i;
   a.href= uri[i];
   a.innerHTML='mylink ' + i;
   iframe.doc.body.appendChild( a );
}
for( var j in iframe.doc.body.childNodes){//cycle trhough links
     var display = getStyle(j, iframe.doc, "display");
     if(display = 'none')
          alert('url ' + j.href + ' has been visited');
}
```

Vari usi:

Groups on SN detection

Blogs comments posting fingerprinting

Gender prediction

EE on it.hackmeeting.org :)

Ci sono tecniche per recuperare informazioni dal vostro browser SENZA sfruttare dati su di esso salvati.

Semplicemente: si raccolgono dati sul vostro sistema.

Come, quali?

User agent
HTTP accept headers
IP address
Plugins
Screen resolution
Clock skew respect to server
JS VM performance
Browser metrics
System Fonts (via flash, java)
Timezone
FF or chrome extensions
Privacy settings
... il limite e' piu' o meno l'immaginazione!

Il fingerprinting si basa innanzitutto sulla capacita' di combinare tutti questi dati finche' non si ottiene un set di dati univoco per il singolo utente.

Esperimento: panopticlick.eff.org

Raccolta di (alcuni) dati sugli utenti, creazione statistiche, determinazione della capacita' di fingerprinting.

Permette di testare quanto e' “univoco” il proprio browser (approccio statistico)

Vi rimando al paper per i dettagli, ma:
- 99+% degli utenti identificabili univocamente

Anche i cambiamenti di fingerprint sono facili da seguire con euristiche banali.

Quindi, per essere “invisibili” bisogna:

disabilitare o mischiare i cookies
randomizzare lo user agent
usare tor e torbutton
disabilitare TUTTE le cache, la history
disabilitare http referer, preferred encoding
disabilitare flash, java, silverlight
usare noscript

E sperate di non esservi scordati nulla!
Fatto questo, cadiamo nel...

PARADOSSO DEL PARANOIDE

Ci sono CONCRETE possibilita' che siate gli unici visitatori di un sito che in quel momento fanno richieste cosi' particolari, quindi siete individuabili

e navighere di merda!

Ma almeno l'anonimato e' salvo!

e questa era la silde 42.

Part III: Conclusions

Quindi:

La profilazione e' pressoche' inevitabile, tanto vale esserne consapevoli, ed evitare di mettersi nelle mani di soluzioni bizzarre o di ciarlatani vari (si', esistono anche soluzioni commerciali per l'anonimato e la resistenza alla profilazione! cfr. abine)

E, in generale, non e' il caso di pensare che un transparent proxy (google sharing) o un'estensione di firefox siano piu' affidabili di google.

Tuttavia ci sono molte situazioni in cui l'anonimato e' molto importante.

Bisogna continuare a lavorare per migliorare il nostro livello di anonimizzazione, creare nuovi strumenti, essere creativi

e cercare di creare strumenti FACILI e diffusi (la diffusione e' la chiave per ottenere l'anonimato).

e soprattutto sosteniamo i server autogestiti. Sono uno spazio di liberta' imprescindibile.

Vi lascio con una citazione

"... solitamente gli anonimi non sono tanto intelligenti ..."

Salvatore Aranzulla

http://aranzulla.tecnologia.virgilio.it/sms-anonimi-dalle-cabine-telefoniche-594.htm