SmoothWall.net Technology White Paper

Google via HTTPS
Risks associated with Google's secure search service

Introduction

May 2010 saw Google announce the beta release of their secure search service - Google via HTTPS. This whitepaper examines the risks search engines may pose and explores how Google's recent changes might challenge filtering products or services.

Search Engines are often regarded as safe locations, but there are several non-trivial concerns for networked environments.

Thumbnails	Objectionable image previews sometimes appear in image search results. Even without access to image search, Google can often return thumbnails as part of a "normal" search process.
Unsafe Results	Without the use of SafeSearch, a plethora of objectionable and even illegal content can be readily located. Even with Safe Search, restrictions are limited to only the most unacceptable content.
Page Caches	Some search engines incorporate cache functionality, which allows their user to access a snapshot of a website from when the search engine examined it. Because this facsimile is served from the search engine's own domain, this can circumvent simple filtering schemes.

Remedies

The issues posed by search engines demand a defence-in-depth approach, involving both the headers and the body of all requests and responses.

Search Term Filtering

By examining outgoing search requests, a web filter can prevent prohibited searches from ever reaching the search engine.

Google via HTTPS
Risks associated with Google's secure search service

Force Safe Search

Most search engines now incorporate safe search options which remove unsavoury items to provide family safe results. A web filter can be used to manipulate search requests, forcing safe search features on regardless of individual clients' settings.

Deep URL Analysis

Deep URL analysis can examine the entirety of a URL string to best identify the true content. This can prevent thumbnails and caches of banned sites being accessed via the "safe" domain of your search engine.

Dynamic Content Analysis™

In the rare cases where forced SafeSearch and banned search terms fail to stop unacceptable search results, Dynamic Content Analysis is employed to analyse the content of all pages in real time, either blocking the results pages or the individual items when they are accessed.

The Impact of the Change

Google now make their web search available via HTTPS. An HTTPS connection will encapsulate the regular HTTP protocol within Secure Sockets Layer (SSL) encryption. This is designed to protect a web connection from eavesdropping on the wire.

The Core Problem

Unfortunately for a web filter, this can negatively impact upon the traditional methods of action for preventing search engine abuse.

A filter that analyses the search terms used on a search engine will examine the HTTP headers of web requests to extract the desired information. However, a secure HTTPS based search will never reveal such information in plain text.

Figure A details an initial request using Google's new SSL search. It is clear to see that a connection is being established to the server 'www.google.com', but from this point onwards all HTTP data is encrypted via SSL and invisible to a simple web filter.

Compare this to Figure B which shows a search for "APPLES" on Google's plain HTTP service; here the exact URL request is available to the filter, along with the full headers. With such information, particular searches can be permitted or denied to good effect.

Figure A

```
CONNECT www.google.com:443 HTTP/1.1
User-Agent: Mozilla/5.0 (Windows; U; Windows
NT 5.1; en-GB;rv:1.9.1.9) Gecko/20100315
Firefox/3.5.9 (.NET CLR 3.5.30729)
Proxy-Connection: keep-alive
Host: www.google.com
```

Figure B

```
GET /search?hl=en&source=hp&q=APPLE&
btnG=Google+Search&aq=f&aqi=&aql=&oq=APPLES&
gs_rfai=&fp=62c7546c9e19fcb7 HTTP/1.0
```

```
Host: www.google.com
User-Agent: Mozilla/5.0 (Windows; U; Windows
NT 5.1; en-GB; rv:1.9.1.9) Gecko/20100315
Firefox/3.5.9 (.NET CLR 3.5.30729)
Accept: text/html,application/
xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Language: en-gb,en;q=0.5
Accept-Encoding: identity,gzip,deflate
Accept-Charset: ISO-8859-1,utf-8;q=0.7,*;q=0.7
Referer: http://www.google.com/webhp?hl=en
Cookie: PREF=ID=aeecf95a1eaed20d:U=d47a532cd5c
2222f:LD=en:NR=10:CR=2:TM=1253288402:LM=126623
8089:GM=1:S=9cYtU6woijAy8eYn;
NID=34=kJR2YEn1BkqUhrGOH3zHlNBymMiKgd_jhs4rOnK
rzjHpDaDSVMEh8hUv_5j7XRd69EG-EoCTty9Qfys
b4xKk4EgS4Q8-ZmZPRLJYdeEUWdUyoCEdRFSfj3zSxJQbf
Y1b;rememberme=false
Via: 1.0 ng08-2:801 (squid/2.7.STABLE6)
Cache-Control: max-age=259200
Connection: keep-alive
```

Solutions

The lack of header data will also cause problems when attempting to force SafeSearch features at the web filter level. Typically these techniques will modify clients' search requests, appending the URL or changing Cookies to set the engine's SafeSearch parameter. An HTTPS connection denies a simple filter the opportunity to identify web searches in progress, since only domain-level information is visible. Furthermore, the SSL encrypted payload cannot be modified, precluding any central control over search options like SafeSearch.

Deep URL analysis suffers from the lack of headers in a similar way.

The final tool in an advanced web filter's box is the dynamic analysis of page content which can occur when the search results are returned. It may be no surprise to the reader that this technique is also ineffective against secure search as the HTTP body data is concealed by SSL encryption until it reaches the browser.

An instinctive and immediate solution, available to most URL filters, is to simply block access to google.com.

Blocking google.com quickly alleviates the problems caused by secure search, since Google have not yet SSL-enabled their localized services on top-level domains (TLDs) such as .co.uk and .fr. Unfortunately, this method prevents clients from reaching Google's regular http service on google.com. This may cause some problems since on the wider web, Google's original US service on .com is linked to regularly. Furthermore 'google.com' dependancies will be found embedded in numerous places, including Google Custom Search widgets on 3rd party websites, browser search toolbars, bookmarks and homepages, and even within software packages. Many such dependencies will be non-trivial, or impossible, to resolve.

Unfortunately for users of Google's many web services, the blocking of google.com may have unwelcome side effects. While most of these services use a dedicated domain (e.g. Gmail – mail.google.com, YouTube – youtube.com), all those services that require a Google accounts login will reach the following secure gateway: https://www.google.com/accounts/ServiceLogin
Recalling Figure A, it is impossible to distinguish between this login gateway and the

secure search page. Both will be blocked, hence logged-in Google services will be unavailable.

The technique of blocking google.com can be finessed by restricting the block strictly to HTTPS requests for https://www.google.com. This will allow searches on the non-secure, HTTP, version of Google to proceed as normal whilst blocking secure search. By doing this, your users will not have to modify their actions, software or bookmarks with respect to regular Google search.

To gain the full benefit of a web filter such as Guardian which can filter search terms, force SafeSearch, monitor Google web cache, and perform Dynamic Content Analysis™ upon results one must utilize SSL Interception.

Summary

SSL Interception ensures that filtering (and Dynamic Content Analysis) is performed on all traffic that utilizes secure or https connections. All of the search engine specific protections will be applied to Google SSL search in the same way as they are to the HTTP version.
SSL Interception is the peace of mind option that extends our comprehensive search engine controls to Google's new secure service. With a growing uptake in institutional use of Google web applications such as Documents, GMail and Calendar, SSL Interception is the clear choice that protects against search engine risks, yet retains the productivity based applications stack.

Without remedial action, the new secure search service from Google provides a successful method of circumventing web filter controls. Undoubtedly this is an issue that must be addressed and the risk will worsen considerably should Google choose to enable Image search, or Universal search on the secure service.

Full interception of SSL connections provides the most complete answer to this issue, allowing Guardian's long-established search engine protections to work with the new Google secure search.

Those without SSL interception stand to lose access to services reliant on Google accounts.

Approaches that cannot perform SSL interception will be severely limited by the inherent features of the protocol. Some of these problems could be remedied in future by the Google account login portal moving to a distinct sub-domain (for example: login.google.com). We understand however, that such a change is of minimal benefit to Google and may not ever occur. Institutions reliant on Google Apps would be well advised to seek an HTTPS capable web filter with quality search engine controls in the very near future. The Guardian range of products has provided such functionality at no additional cost for a number of years.