Google Black Hat Sitemap Bug: What It Means for XML Sitemaps by @TomAnthonySEO

A few months back I discovered a shocking bug in how Google handles XML sitemaps, which enabled brand new sites to rank for competitive shopping terms by hijacking the equity from legitimate sites.

I reported this issue to Google, and they have now fixed the issue, and paid me a bug bounty.

However, since I published my write up of the issue, a number of SEO professionals have contacted me worrying that they may have been a victim of such an attack, requesting that I help use the attack or theorizing variations which may still work.

This article will answer some of the most popular questions I've been getting.

What Was Google's XML Sitemap Bug?

The issue is related to how Google handles and authenticates XML sitemap files, specifically those files that were submitted via the ping mechanism.

Sitemaps can be submitted directly to Google Search Console, via an entry in your robots.txt file, or by 'pinging' them by sending the sitemap URL to a special endpoint that Google provides.

For GSC and robots.txt entries these are obviously authenticated as genuine by the fact you have entry to the

domain's GSC or robots.txt file, but for ping URLs, Google seemed to decide whether they were trustworthy simply by looking at the domain in the URL that you send.

The issue is that if this URL redirects elsewhere, even to a different domain, then Google still trusts it as belonging to that original URL.

So, for example, I may submit a sitemap URL of apples.com/sitemap.xml, but that URL may redirect to oranges.com/sitemap.xml, but Google would still associate the XML sitemap as belonging to apples.com.

What Are Open Redirects?

Many websites succumb to a form of manipulation known as "open redirects," where an attacker can trick a website into redirecting to a location of their choice.

An example may be websites that have a login mechanism that has the form apples.com/login.php?continue=/shop, which may be manipulated to be apples.com/logout.php?continue=http://evil.com/.

In my research, I found open redirects on Facebook, LinkedIn, Tesco, and a number of other sites (I've reported all of these, and many have been fixed).

To give an indication of how widespread they are, Google's Vulnerability Rewards Program explicitly excludes open redirects as qualifying for a bounty (and indeed there are known open redirects on Google).

This allowed for the opportunity to ping sitemaps via an open redirect on a legitimate site which would redirect to the XML file hosted on an attacker site.

For example, by submitting a sitemap on the URL apples.com/logout.php?continue=http://evil.com/sitemap.xml,

Google would treat it as being an authentic sitemap for apples.com, but it would actually be hosted on evil.com.

At this point, evil.com can submit sitemaps for apples.com, and by including hreflang entries in these sitemaps, it can leverage apples.com's equity (PageRank) to rank for search results it has no legitimate right to do so.

Are You a Victim & Now Being Outranked?

Since the news became public, I have had a bunch of SEO professionals from various places reach out to me asking me to review their case, concerned that they may have been the victim of this or asking if this is how a competitor is able to outrank them.

I can certainly understand why.

It can sometimes be super frustrating to try to understand why another site is ranking so well against you, or why your site has suddenly had a lull in performance.

Having an explanation for these edge cases is certainly appealing.

So far I have not seen anything to convince me that this bug was being exploited in the wild.

Google is a complex beast, and there could be all sorts of explanations for why certain sites are ranking the way they are, but at the moment I remain to be convinced that this bug is one of them.

If you are concerned you are the victim of this, then the only real footprint it would leave is an entry in your server logs showing Googlebot coming to your site to collect a sitemap and being 3xx redirected to another domain (JavaScript and meta-

refresh redirects wouldn’t work).

This is the best thing you can check.

In my experiment, I was regularly re-pinging the sitemap, but even without re-pings it I believe Google would always go via the open redirect, so you should see entries in your server logs.

Does This Change Anything About XML Sitemaps?

Yes. It changes when hreflang entries will be used.

Google will no longer pay attention to hreflang entries in “unverified sitemaps”, which I believe means those submitted via the ping URL.

Those submitted inside Google Search Console or in your robots.txt file will still operate as they always have done, and pinging one of these sitemaps to prompt a recrawl from Google will also work as expected.

I anticipate the change will affect very few sites, but you should be aware of it.

Conclusion

My recommendation: submit sitemaps via both the GSC interface and include them inside your robots.txt.

If you are a site that suffers particularly from scrapers, for whatever reason, then you may wish to exclude sitemap entries from your robots.txt file such that bad actors cannot find them and use them to expedite their efforts.

More XML Sitemap Resources:

http://tracking.feedpress.it/link/13962/8895432