GSB output Analysis – site pages

(as contrary to AS pages)

Table of Contents

Google Safe Browsing – an Analysis for Site HTML files

Intro

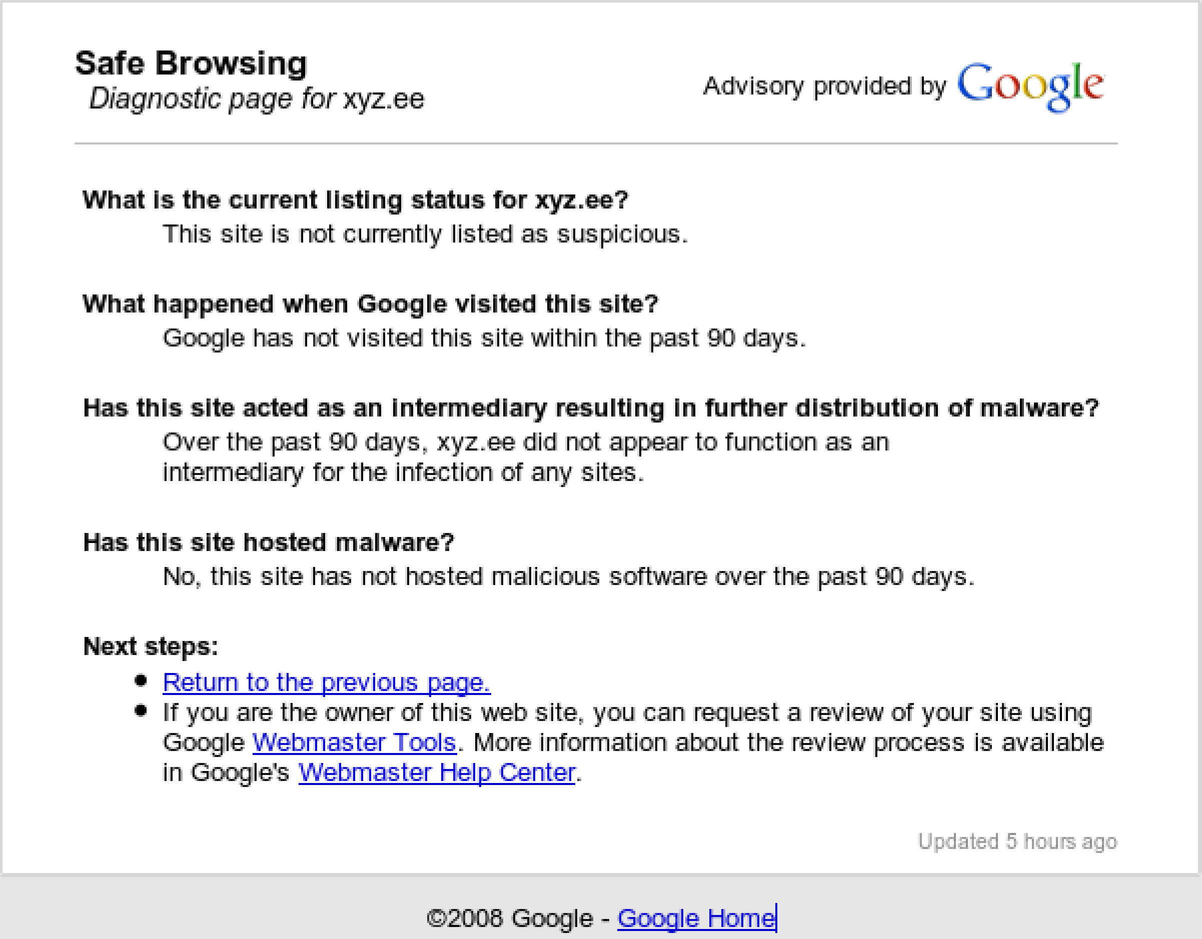
We are discussing the FQDN (site, domain) type HTML files downloaded from Google SafeBrowsing. (AS related files have a slightly different structure and will be possibly described elsewhere.)

It is possible to parse both ways – as HTML or text but html parsing seems to be more reliable. Disclaimer – we only are considering English pages.

General Examples

URL to obtain: http://safebrowsing.clients.google.com/safebrowsing/diagnostic?site=xyz.ee

Example output of a clean site:



Safe Browsing
Diagnostic page for xyz.ee

Advisory provided by Google

What is the current listing status for xyz.ee?
This site is not currently listed as suspicious.

What happened when Google visited this site?
Google has not visited this site within the past 90 days.

Has this site acted as an intermediary resulting in further distribution of malware?
Over the past 90 days, xyz.ee did not appear to function as an intermediary for the infection of any sites.

Has this site hosted malware?
No, this site has not hosted malicious software over the past 90 days.

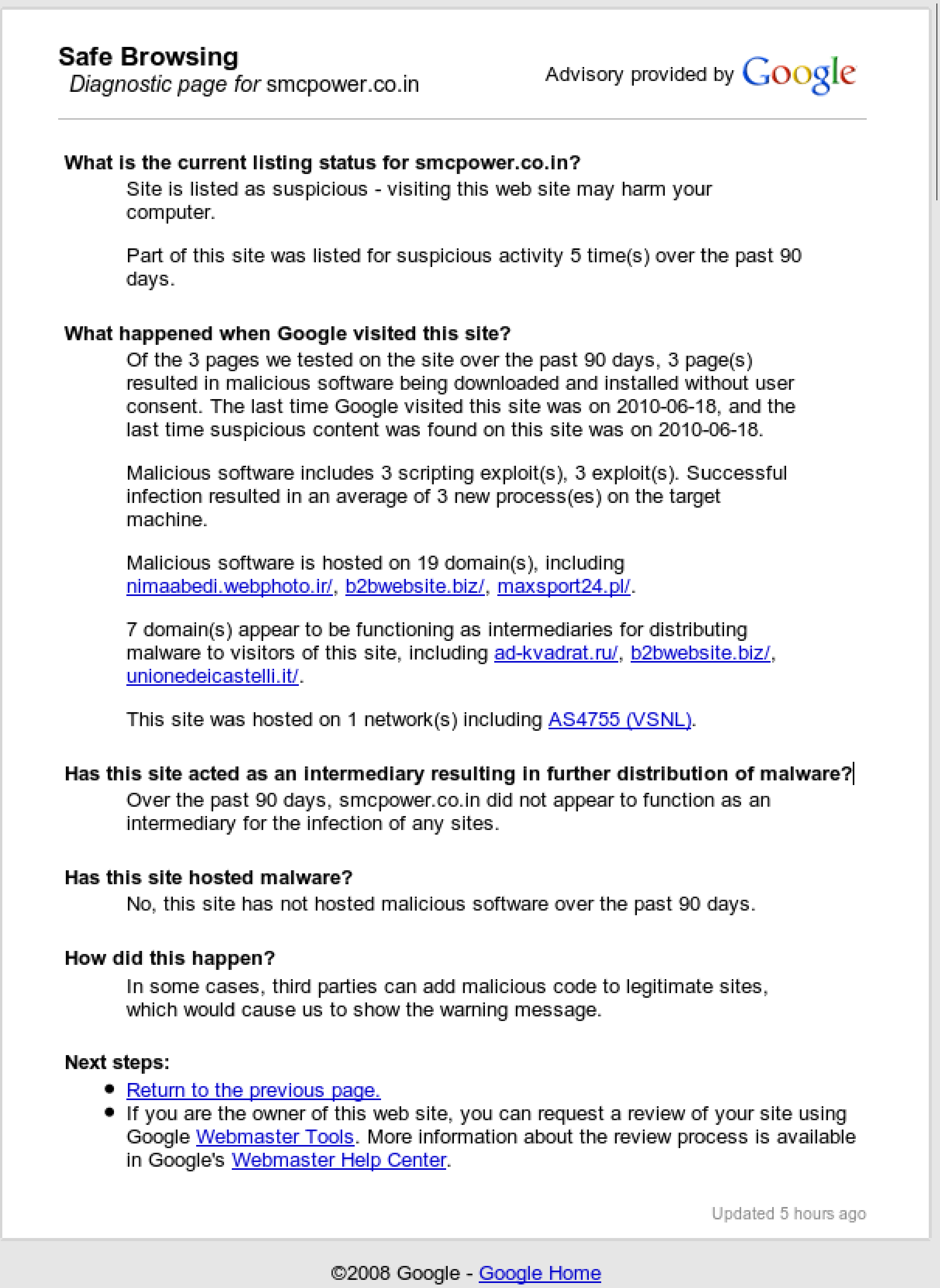
Next steps:

- Return to the previous page.
- If you are the owner of this web site, you can request a review of your site using Google Webmaster Tools. More information about the review process is available in Google's Webmaster Help Center.

Updated 5 hours ago

©2008 Google - Google Home

Example for a rather dirty site:



Safe Browsing
Diagnostic page for smcpower.co.in

Advisory provided by Google

What is the current listing status for smcpower.co.in?

Site is listed as suspicious - visiting this web site may harm your computer.

Part of this site was listed for suspicious activity 5 time(s) over the past 90 days.

What happened when Google visited this site?

Of the 3 pages we tested on the site over the past 90 days, 3 page(s) resulted in malicious software being downloaded and installed without user consent. The last time Google visited this site was on 2010-06-18, and the last time suspicious content was found on this site was on 2010-06-18.

Malicious software includes 3 scripting exploit(s), 3 exploit(s). Successful infection resulted in an average of 3 new process(es) on the target machine.

Malicious software is hosted on 19 domain(s), including nimaabedi.webphoto.ir/, b2bwebsite.biz/, maxsport24.pl/.

7 domain(s) appear to be functioning as intermediaries for distributing malware to visitors of this site, including ad-kvadrat.ru/, b2bwebsite.biz/, unionedeicastelli.it/.

This site was hosted on 1 network(s) including AS4755 (VSNL).

Has this site acted as an intermediary resulting in further distribution of malware?

Over the past 90 days, smcpower.co.in did not appear to function as an intermediary for the infection of any sites.

Has this site hosted malware?

No, this site has not hosted malicious software over the past 90 days.

How did this happen?

In some cases, third parties can add malicious code to legitimate sites, which would cause us to show the warning message.

Next steps:

- Return to the previous page.
- If you are the owner of this web site, you can request a review of your site using Google Webmaster Tools. More information about the review process is available in Google's Webmaster Help Center.

Updated 5 hours ago

©2008 Google - Google Home

Page layout

Within the HTML page, overall 8 sections could be defined of which only 6 are useful (except F,G). Here is the ontology:

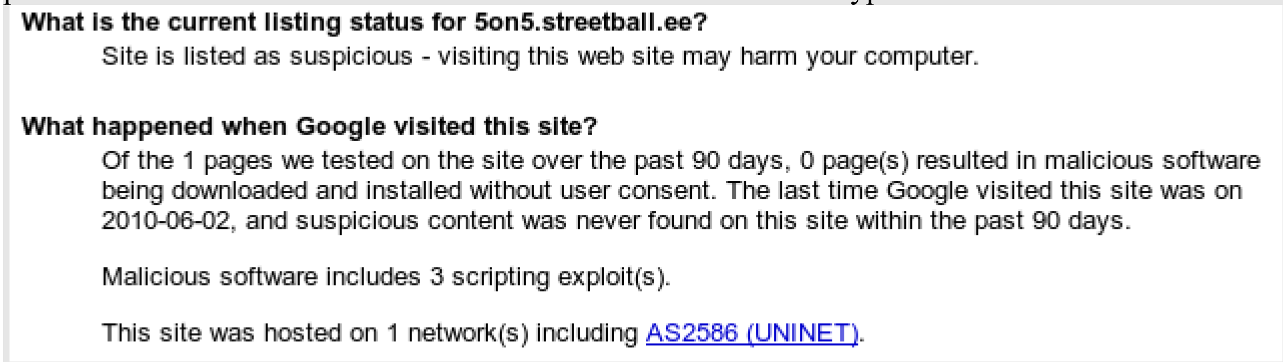
- A - Static Page header
- B – Listed(ness) Status - „**What is the current listing status for**“
- C- Visit results - „**What happened when Google visited this site?**“ (the most informative section)
- D – Intermediary - „**Has this site acted as an intermediary resulting in further distribution of malware?**“
- E - Hosting - „**Has this site hosted malware?**“
- F – Rhetorical - „**How did this happen?**“
- G – Useless - „**Next steps:**“ (oriented to dumb-users)
- H – Footer including the estimation of data age - „Updated 5 hours ago“

Artefacts

It should be noted that the Google data quality is not perfect. Some known examples:

* Estonian translation for the listedness status is wrong – all sites are „suspicious“

* Internally contradicting data presented – **malware never found** but containing **3 exploits**. It is possible that some fine intrinsices exist in malware classification types:



What is the current listing status for 5on5.streetball.ee?

Site is listed as suspicious - visiting this web site may harm your computer.

What happened when Google visited this site?

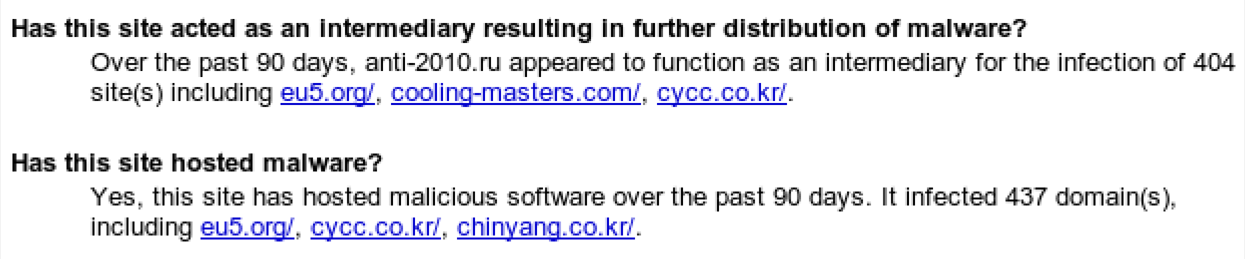
Of the 1 pages we tested on the site over the past 90 days, 0 page(s) resulted in malicious software being downloaded and installed without user consent. The last time Google visited this site was on 2010-06-02, and suspicious content was never found on this site within the past 90 days.

Malicious software includes 3 scripting exploit(s).

This site was hosted on 1 network(s) including AS2586 (UNINET).

* The semantics or Sections C, D and E is not fully clear. What is the role of intermediary while a normal 0-day virus usually goes via 4-5 sequential hops before actually downloading the payload.

* Sometimes a particular site is listed as both victim and offender (note **eu5.org** and **cycc.co.kr** in both categories):



Has this site acted as an intermediary resulting in further distribution of malware?

Over the past 90 days, anti-2010.ru appeared to function as an intermediary for the infection of 404 site(s) including eu5.org/, cooling-masters.com/, cycc.co.kr/.

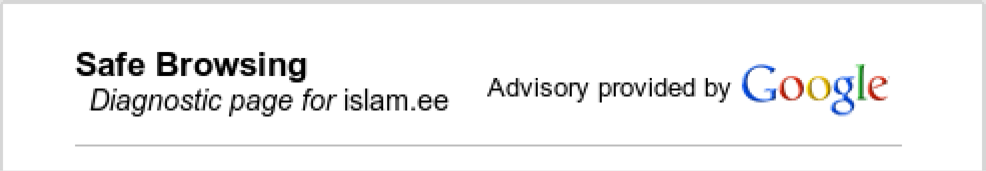
Has this site hosted malware?

Yes, this site has hosted malicious software over the past 90 days. It infected 437 domain(s), including eu5.org/, cycc.co.kr/, chinyang.co.kr/.

- „**90 days**“ is a constant – it never changes. We have seen „**hours ago**“ but never „days ago“.

Sections

A - Page Header



Safe Browsing
Diagnostic page for islam.ee

Advisory provided by Google

Q1 – Obtain the sitename -> **islam.ee**
Alternatively it could be obtained from page title (HTML).

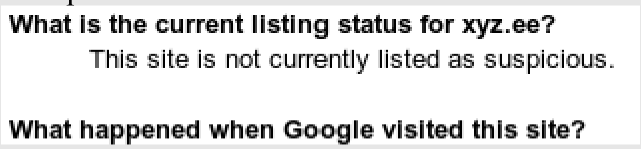
Requests against top-level domains are not resulting any sensible GSB answer. Generally, when many subdomains are present, then, for situation awareness, some separate requests should be provided, e.g each for for:

- www.staff.ttu.ee
- staff.ttu.ee
- ttu.ee

B – Listed Status

Examples

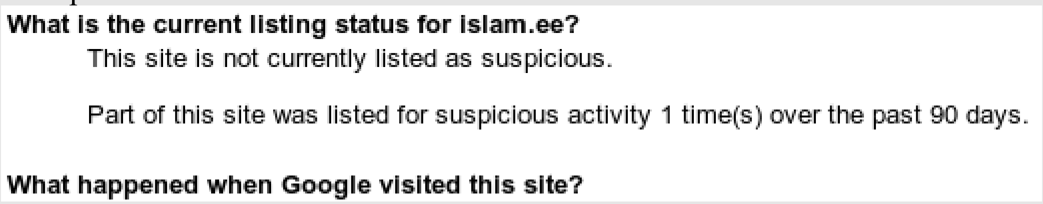
Example for site B1:



What is the current listing status for xyz.ee?
This site is not currently listed as suspicious.

What happened when Google visited this site?

Example for site B2:

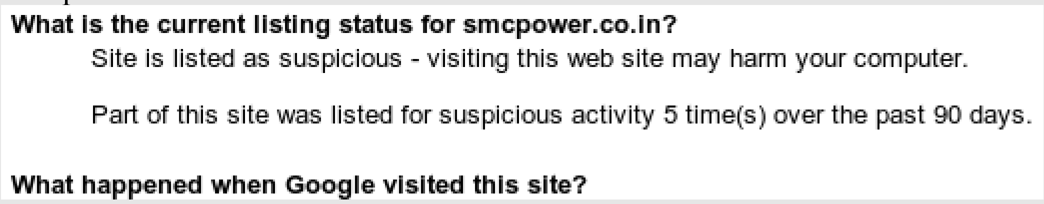


What is the current listing status for islam.ee?
This site is not currently listed as suspicious.

Part of this site was listed for suspicious activity 1 time(s) over the past 90 days.

What happened when Google visited this site?

Example for site B3:



What is the current listing status for smcpower.co.in?
Site is listed as suspicious - visiting this web site may harm your computer.

Part of this site was listed for suspicious activity 5 time(s) over the past 90 days.

What happened when Google visited this site?

Parsing

Parsing Section B is straightforward:

Q1 – is the site listed or not. That section is always present and the answer is one of two:

- Site is listed as suspicious - visiting this web site may harm your computer.
- This site is not currently listed as suspicious.

Q2 – whether the second statement is present at all - „**Part of this site was listed for suspicious activity 5 time(s) over the past 90 days.**“

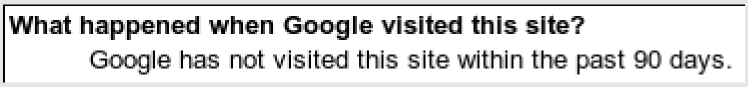
Q3 – if it is present, the NUMBER between words „**activity**“ and „**time(s)**“ could be obtained.

Section C – Visit Results

This is absolutely the most complex and critical section. It is most difficult to parse due to the fact that certain subsections could be present or missing at Googles will.

Examples

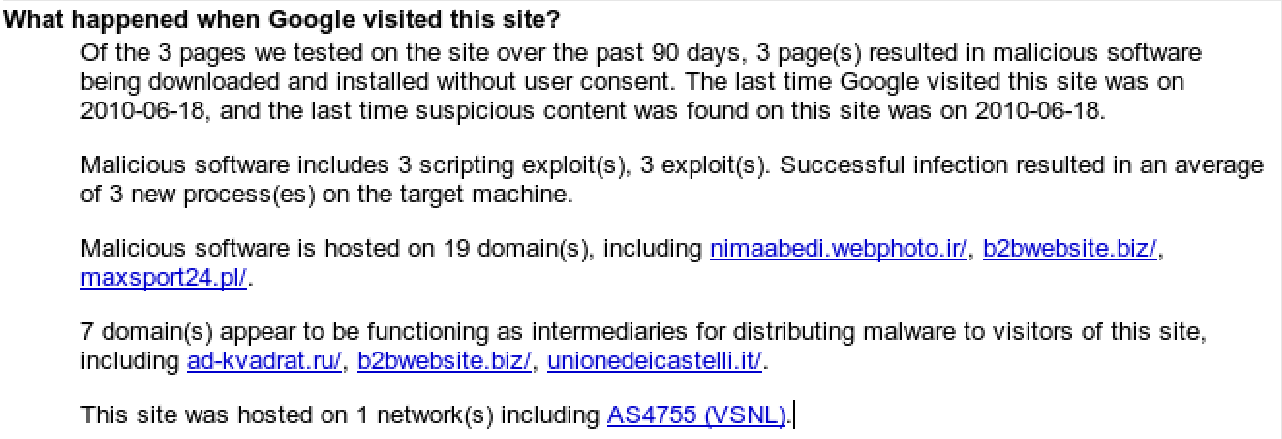
Example of a site never visited.



What happened when Google visited this site?

Google has not visited this site within the past 90 days.

Example of a visited site:



What happened when Google visited this site?

Of the 3 pages we tested on the site over the past 90 days, 3 page(s) resulted in malicious software being downloaded and installed without user consent. The last time Google visited this site was on 2010-06-18, and the last time suspicious content was found on this site was on 2010-06-18.

Malicious software includes 3 scripting exploit(s), 3 exploit(s). Successful infection resulted in an average of 3 new process(es) on the target machine.

Malicious software is hosted on 19 domain(s), including nimaabedi.webphoto.ir/, b2bwebsite.biz/, maxsport24.pl/.

7 domain(s) appear to be functioning as intermediaries for distributing malware to visitors of this site, including ad-kvadrat.ru/, b2bwebsite.biz/, unionedeicastelli.it/.

This site was hosted on 1 network(s) including AS4755 (VSNL).

The difficulty is that subsections C-3 and C-4 both start with „Malicious software“. Lists of URLs (if present at all) could have 1-3 elements. Keywords „domain(s)“, „exploit(s)“ never change the quantity (singular/plural) which makes parsing easier.

The layout of Section C

Here is the ontology:

- Imaginary Section C0 – it defines whether at all to parse Section C in deep
- Section C1 – provides pagecount
- Section C2 - provides ISO dates
- Section C3 – lists exploit metrics
- Section C4 – counts the subdomains, lists up to 3 of them
- Section C5 – counts intermediary domains, up to 3 (we do not have the exact definition)
- Section C6 – counts AS connectivity and lists up to 3 of these with numbers and names

Parsing Section C0

Answer to that question defines whether the site is visited or not, according to the two examples above.

Q1 – Whether the string „Google has not visited this site within the past 90 days.“ is present?
If yes, then establish Has_Visited_Last_90_Days and don't care about the rest of the section.

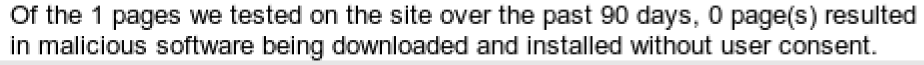
Parsing Section C1

There is a need to further parse section C only if the site is a visited one (i.e. no C0 applies).
NB! Sections C1 and C2 are not visually independent, however, the separation is better at HTML level.

This section will result in two integers:

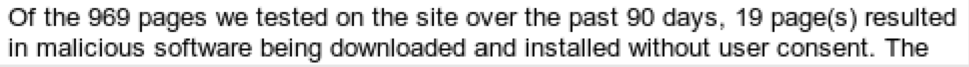
Q1- how many pages they tested on the site – 1 or more.
Q2 - how many pages were found dirty – 0 or more.

Example with a clean result



Of the 1 pages we tested on the site over the past 90 days, 0 page(s) resulted
in malicious software being downloaded and installed without user consent.

Example with a dirty result:



Of the 969 pages we tested on the site over the past 90 days, 19 page(s) resulted
in malicious software being downloaded and installed without user consent. The

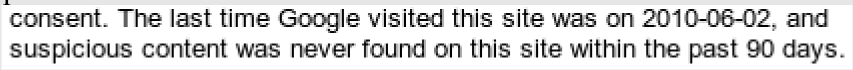
NB! The dirtyness established in this subsection (and possibly kept in a variable) is not fully defining the content of the the following sections. Due to the Google quirks, one must attempt a full parsing even if the site is claimed „not dirty“ here.

Parsing Section C2

This section will result one **or** two ISO dates (**YYYY-MM-DD**) and possibly some meta variables.

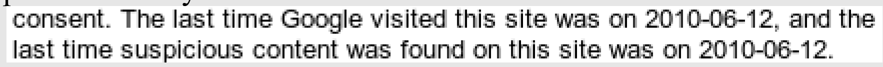
NB! Sections C1 and C2 are not visually independent but logically.

Example with a clean result:



consent. The last time Google visited this site was on 2010-06-02, and suspicious content was never found on this site within the past 90 days.

Example with a dirty result:



consent. The last time Google visited this site was on 2010-06-12, and the last time suspicious content was found on this site was on 2010-06-12.

Q1 – Last visit date is obtainedtween the words „...this site was on“ and „, and the last time...“
Q2 – Need to establish whether another ISO date exist at all or is it replaced to a common „90 days“ phrase. Literally, this is the variable Shit_Ever_Found.
Q3 – If it exist, another ISO date value should be taken after the words „...was found on this site was on „ but the end marker is extremely difficult to place except when using HTML.

NB! However, due to Google inconsistencies, they could provide some unexpected data in other sections, despite the artefact that „suspicious content was never found“.

Further date magic.

There are some hidden data relations relevant to datevalues:
Q4 – Whether **LastVisited** and **Dirty** Data Values are the same? If there are, we probably have an alarming condition and want to raise an alarm. Literally the variable could be **ShouldRaiseAlarm**.

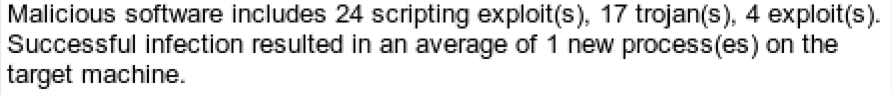
However, even more date magic is possible. It is possible to establish a relationship between dates here and the age value in section H. This way we could estimate the freshness of the information.

Obtaining Google HTML files and parsing these also takes some times. If needed, the file creation dates could further be taken into account or a simplification made that the parsing is „near instant“.

Parsing Section C3

The section is optional and, if missing, easily confused with the next one. The section results some metrics regarding the malware types available on the site.

Example:



Malicious software includes 24 scripting exploit(s), 17 trojan(s), 4 exploit(s). Successful infection resulted in an average of 1 new process(es) on the target machine.

There are difficulties to provide a **full** list because all types are not present on one site.

The list of parsable elements, mostly known types of malware (possibly an incomplete list):
130 worm(s), 77 scripting exploit(s), 10 exploit(s), 35 trojan(s), 2 bot(s). Probably **backdoors** and **viruses** are missing from our list.

Additionally, the **average new process count** is calculated in accordance with **some unknown kind** of the infection (have seen the count as high as 8). That figure is optional, i.e. often not produced.

These integers should be carefully parsed out of the text. Here is the typology:

Q1 – Number of worms, if any
Q2 – Number of scripting exploits, if any
Q3 – Number of exploits, if any
Q4 – Number of trojans, if any
Q5 – Number of bots, if any
Q6 – Something else, was it Number of Backdoors? Have seen earlier.
Q7 – Something else, was it Number of Viruses? Have seen it earlier.
Q8 – (Average) Number of New Processes as the result of Infection, if any.

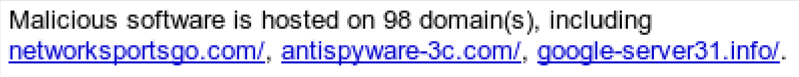
Each number should be parsed separately, there is no certainty in sequence or availability of anyone. There is even no certainty that the parsing instructions here are complete.

Parsing Section C4

The subsection, if present at all, enumerates the domains the badness has been located. Sometimes these are subdomains, sometimes not at all.

The section will result one integer and 1-3 domain names. The section is easy to confuse with the previous one (also starting with „Malicious software...“).

Example:



Malicious software is hosted on 98 domain(s), including networksportsgo.com/, antispyware-3c.com/, google-server31.info/.

NB! 1-3 sitenames are provided, there seem to be normalized (e.g. no **x.co.uk** and **y.co.uk** nearby)

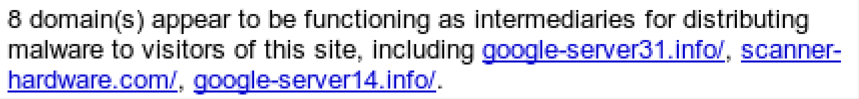
Q1 – Whether the section is present.
Q2 – Obtain the integer from between „is hosted on“ and „domain(s),“
Q3 – Obtain first sitename if any
Q4 – Obtain next (second) sitename, if any
Q5 – Obtain last (third) sitename if any.

Note – there is absolutely no need to preserve URL markup because it mirrors the site names fully.

Parsing Section C5

The section, if present at all, will result the number of intermediaties and possibly list 1-3 bad sitenames.

Example:



8 domain(s) appear to be functioning as intermediaries for distributing malware to visitors of this site, including google-server31.info/, scanner-hardware.com/, google-server14.info/.

The section is a little bit difficult to parse because no start marker present (except HTML).

Q1 – Whether this section is present at all?
Q2 – Obtain the integer before the keywords „domain(s) appear“.
Q3 – Obtain first sitename if any
Q4 – Obtain next (second) sitename, if any
Q5 – Obtain last (third) sitename if any.

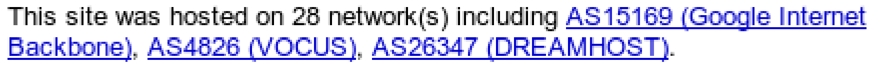
Note – there is absolutely no need to preserve URL markup because it mirrors the site names fully.

NB! There is a confusion with section D. The definition of „**intermediate**“ is unclear. Section D has its own intermediates. So far we haven't seen an example with both sections C4 and D present.

Parsing Section C6

Provided the site has been visited at all, this section always results at least one AS number(s) and name(s) on which the site resides. It also results the integer counting the ASs which site is connected to. However, even if ASN_Count >3 then only 3 ASs are listed.

Example:



This site was hosted on 28 network(s) including AS15169 (Google Internet Backbone), AS4826 (VOCUS), AS26347 (DREAMHOST).

Q1 – Number of networks. Parsed out from between „was hosted on “ and „networks(s) including “.
Q2 – ASN for the first AS, if any
Q3 – AS Name between „(„ and „)“ for that ASN
Q4 – ASN for the second AS, if any
Q5 – AS Name for that ASN
Q6 – ASN for the third AS, if any
Q7 – AS Name for that ASN

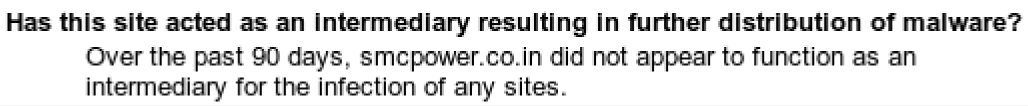
Subsection C6 is the concluding one in section C

D – Intermediaries

NB! This section possibly conflicts with the Section C4 which also deals with intermediaries.

Section D is always present. It either has the „dummy“ content (let's call it D0)

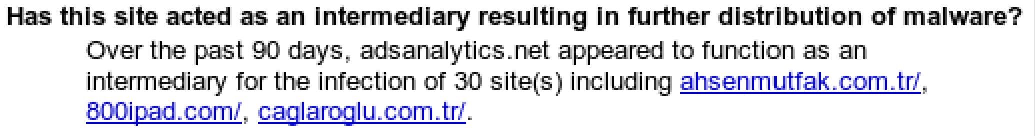
Example – no intermediaries known



Has this site acted as an intermediary resulting in further distribution of malware?

Over the past 90 days, smcpower.co.in did not appear to function as an intermediary for the infection of any sites.

or there are some intermediaries known in which case the section should be parsed:



Has this site acted as an intermediary resulting in further distribution of malware?

Over the past 90 days, adsanalytics.net appeared to function as an intermediary for the infection of 30 site(s) including ahsenmutfak.com.tr/, 800ipad.com/, caglaroglu.com.tr/.

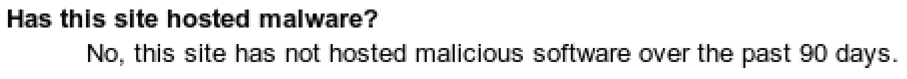
Q1 – whether the dummy Subsection D0 is present. If yes, then no need to parse the section any further.
Q2 – Obtain the numberic between words „for the infection of“ and „site(s) including“.
Q3 – Obtain first sitename if any
Q4 – Obtain next (second) sitename, if any
Q5 – Obtain last (third) sitename if any.

Note – there is no need to preserve URL markup because it is identical to the site names.

E – Hosting

The section explains whether the site itself has hosted malware (whatever it means). If yes, the number of successfully infected victims will be produced and up to 3 (site)names of these. The section seems to be mandatory.

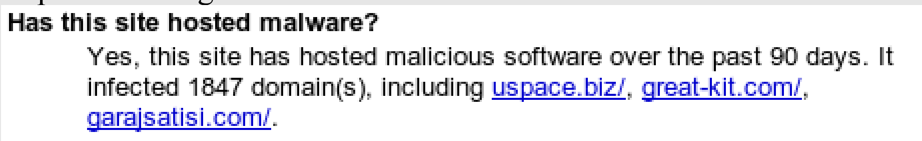
Example of nonhosting:



Has this site hosted malware?

No, this site has not hosted malicious software over the past 90 days.

Example of hosting:



Has this site hosted malware?

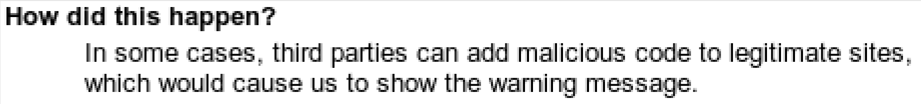
Yes, this site has hosted malicious software over the past 90 days. It infected 1847 domain(s), including uspace.biz/, great-kit.com/, garajsatisi.com/.

Q1 – whether the dummy subsection E0 is present - nothosted. If yes, then no need to parse the section any further.
Q2 – Obtain the numberic between words „It infected“ and „domain(s) including“.
Q3 – Obtain first sitename if any
Q4 – Obtain next (second) sitename, if any
Q5 – Obtain last (third) sitename if any.

Note – there is no need to preserve URL markup because it is identical to the site names.

F – Rhetorical

Section F is useless for the purpose. Morover, the section is optional and the condition is not clear.
Example:



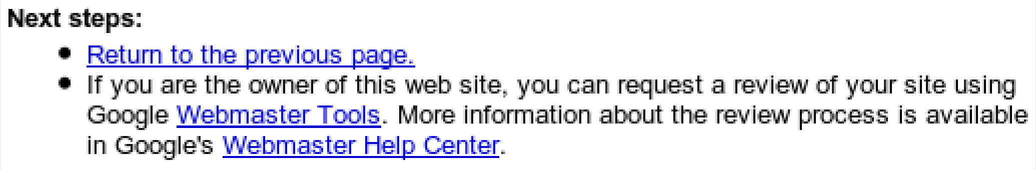
How did this happen?

In some cases, third parties can add malicious code to legitimate sites, which would cause us to show the warning message.

G – Next Steps

Section G is always present but useless for the purpose.

Example:



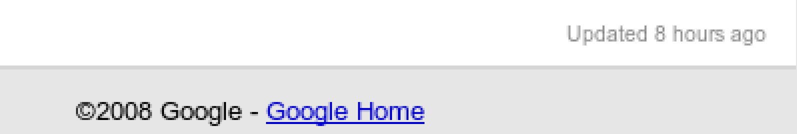
Next steps:

- Return to the previous page.
- If you are the owner of this web site, you can request a review of your site using Google Webmaster Tools. More information about the review process is available in Google's Webmaster Help Center.

H – Footer

Section H has only one interesting attribute which is the „time ago“ value Google uses to estimate the freshness of the data. That value could be used to calculate the lag or to be substracted from the parsing time (time diff between Google updated the data and you retrieved it). In reality, most pages are updated once per 24h and the only value of that number is to estimate, WHEN exactly.

Example:



Updated 8 hours ago

©2008 Google - Google Home

Q1 – Obtain time lag numeric. It lays between words „Updated“ and „hours ago“.

H was the last section which concludes the parsing process.

2010-06-30