YGN ETHICAL HACKER GROUP

Whitepaper

Disclosure Vulnerability: robots.txt

By

d0ubl3_h3lix

Sun Jul 16, 2006

Robots.txt file is used to stop certain web crawlers no matter what they are browsers or search bots like Google-bot. It is placed at root folder. An example goes as:

```
 # to stop IE browsers/users from viewing your site
User-Agent: MSIECrawler
Disallow:/path/
# to stop google spider from crawing your site
User-Agent: googlebot
Disallow:/path/
```

From the security standpoint, robots.txt can be viewed by everyone.
Just type in the browser address bar and go:

```
http://www.somesite.com/robots.txt
```

So, bad guys can gain sensitive information if you store it in robots.txt.

Normally which kinds of folder you don't want bots not to crawl and index?
The likely point is your personal folders, or just your temporary half-finished development stuffs, or kind of those things alike.

Depending on your needs or situation, you're enabling read access to those folders. Then, everyone can steal your assets. If the web site is company site, then there may have sensitive info more or less in robots.txt file.

Countermeasures:

Don't put any of sensitive data on the servers and in robots.txt.
If you want to track attackers who attempt to access your robots.txt, use Google Honeypot. Search it in http://sf.net.

Group Contributions:

Added by br0 at Jan 05th, 2008:

We can find that robots.txt easily by Google hacking technique.
Type

```
intitle:index of robots.txt
```

in Google search box and you may see many sites with robots.txt on search results .
After that you can try to type http://yoursite.com/robot.txt in your browser or click on search result and you can see what inside of robots.txt file in browser.