

robots.txt

Ben Allen

9 Nov 2006

UNIVERSITY OF MINNESOTA

robots.txt

- What is it? Why have one?
- Risks
- Alternatives
- Statistics
- Summary

What is robots.txt ?

- a mechanism to control what parts of your site ***well-behaved*** robots will index.
- a text file placed on a web server
- a list of possible places to look for content that's not indexed
- example:

```
User-agent: *  
Disallow: /admin  
Disallow: /homework-solutions  
Disallow: /grades
```

Why have robots.txt ?

- There *are* well-behaved robots/spiders/crawlers
- Prevent indexing by search engines
- Listen for “poorly” behaving clients
- Prove to search engines (google) that you maintain the site when requesting to have content removed

- <http://www.google.com/support/webmasters/bin/answer.py?answer=35303>

Risks

- Divulges exactly those things you do NOT want people to look at
- False sense of security
- **DOES NOT PROTECT DATA!**

Alternatives

For legally protected data, the best alternative is

**DO NOT PUT PROTECTED
DATA ON A WEBSERVER!**

Alternatives

- If legally protected data **MUST** be placed on a web server:
 - Notify OITSEC `abuse AT umn DOT edu`
 - use **ONLY** HTTPS
 - use ACLs:
 - user & pass required (HTTP Basic, CookieAuth, NTLM, etc.)
 - limit IPs that can access (Allow/Deny, Host Firewall)
 - log access, review logs

Alternatives

For data which is NOT legally protected:

- use ACLs
 - user & pass required (HTTP Basic, CookieAuth, NTLM, etc.)
 - limit IPs that can access (Allow/Deny, Host Firewall)
- use robots.txt for root dir only
- virtual / dedicated host

```
User-agent: *  
Disallow: /
```

Question: If I do not want this to be found by a search engine, why did I place it online?

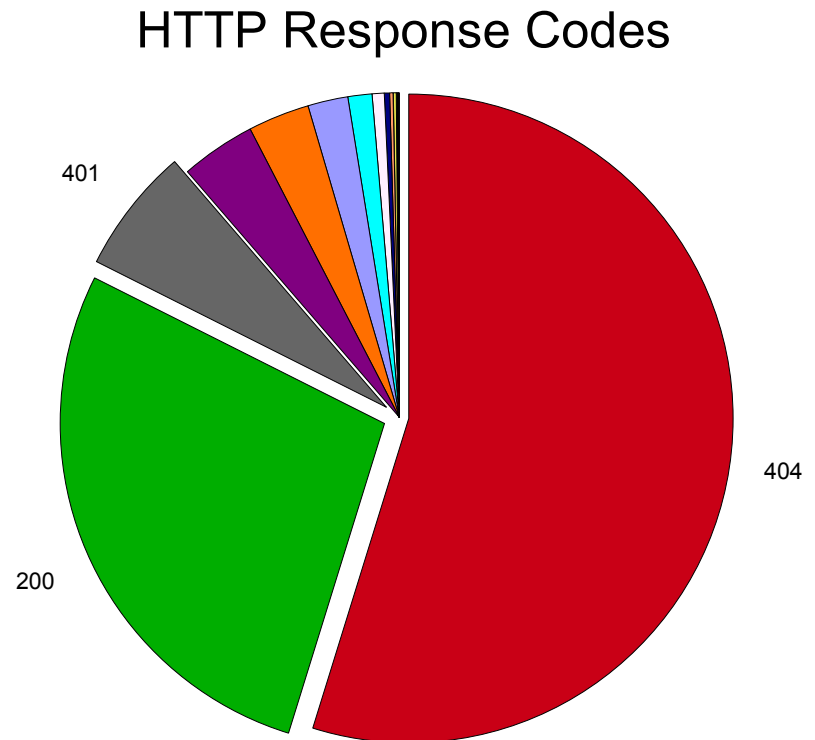
Statistics

- Scan of UMN hosts listening on TCP/80, request `/robots.txt` from the web server.
- Record HTTP headers, server response
- What can we find??

Statistics – HTTP Response

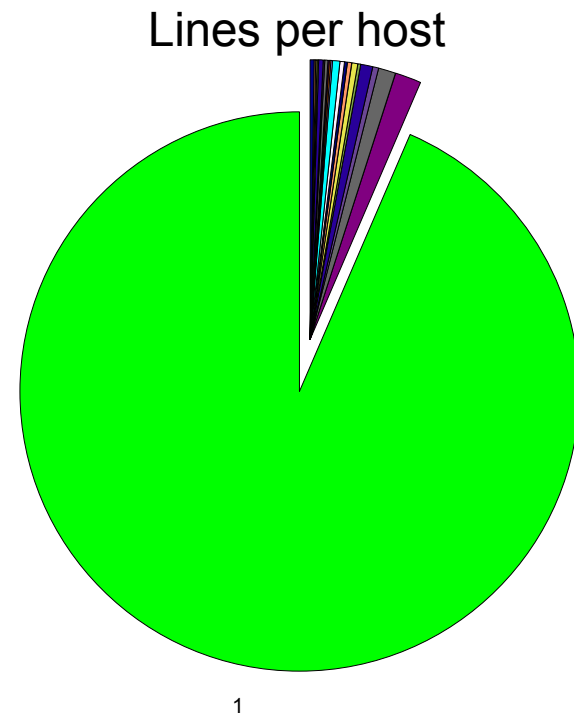
HTTP Responses

- 5119 hosts tcp/80
- 4572 returned headers Responses
- 404(Not Found): 2497
- 200(OK): 1264
- 401(Unauth'd): 292



Statistics - Disallow

- Unique Dirs: 519
- plain “/”: 1103
- max lines: 44
- only one line: 1106



Statistics – Curious Lines

```
Disallow: /abnormal/  
Disallow: /administrator/  
Disallow: /admissions/committee/manual  
Disallow: /backup/  
Disallow: /~joeuser/5573.Fall.2004/  
Disallow: /~joeuser/5674.Spring.2005/  
Disallow: /~joeuser/5162.Fall.2004/  
Disallow: /Configuration/  
Disallow: /*.doc$  
Disallow: /*.xls$  
Disallow: /*.ppt$  
Disallow: /Do_not_delete  
Disallow: /Employees/  
Disallow: /filesNotInUse/  
Disallow: /hr/  
Disallow: /installation/  
Disallow: /my/research/extreme/docs/experiment_consent.pdf  
Disallow: /movies/  
Disallow: /risk_summary/  
Disallow: /sysadmin/  
Disallow: /timecard/  
Disallow: /Tools & Examples/  
Disallow: /vnc/  
Disallow: /_vti_inf.html  
Disallow: /_vti_log/
```

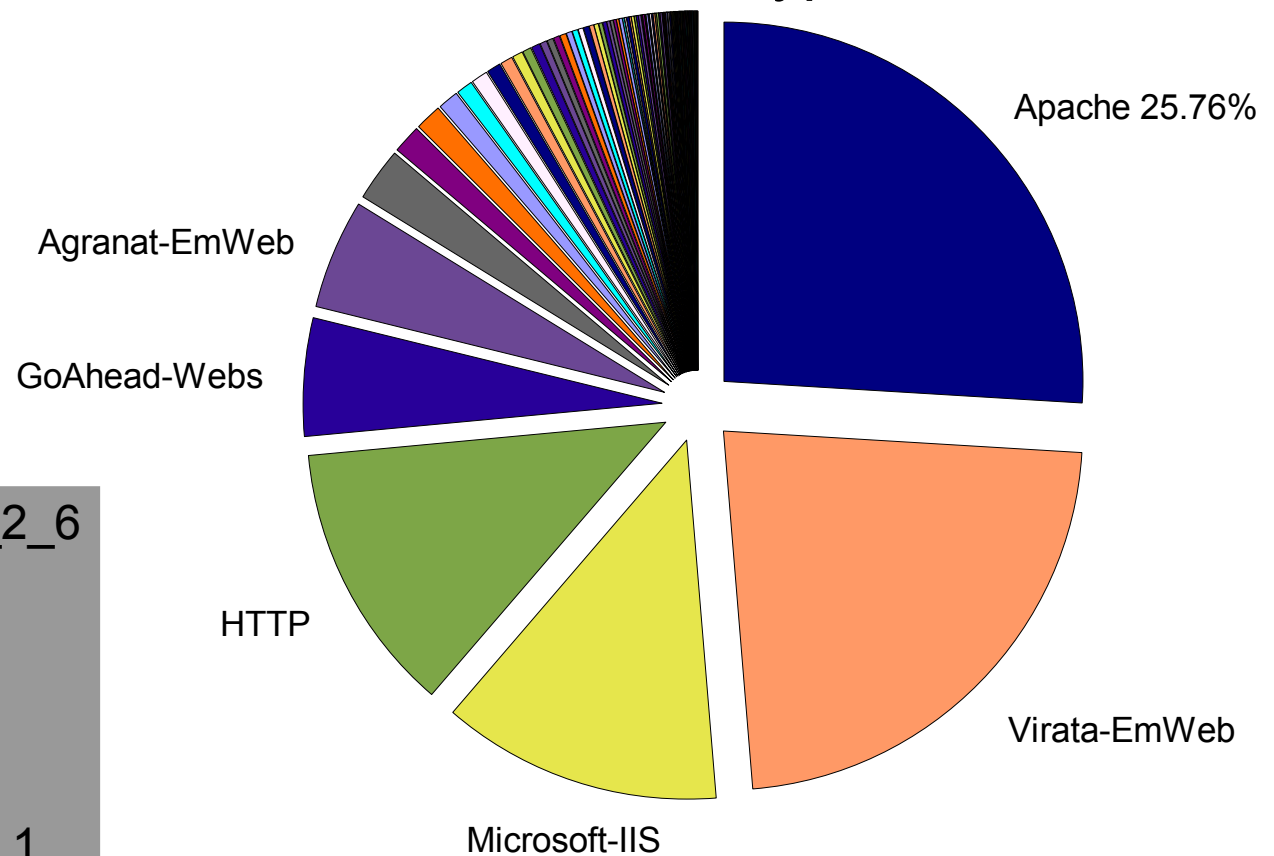
Statistics – User-agent

```
1 User-agent: curl/7.15.5
1 User-agent: ExtensionInternal
1 User-agent: Googlebot
1 User-agent: Googlebot-IA
1 User-agent: Seeker
1 User-agent: Slurp
1 User-agent: Yahoo-Blogs/v3.9
1 User-agent: Yahoo-MMCrawler
1 User-agent: YahooSeeker
1 User-agent: Yahoo-Seeker
2 User-agent: CIDRAP-crawler
3 User-agent: googlebot
3 User-agent: Roverbot
12 User-agent:
1170 User-agent: *
```

Statistics – Web Servers

- Server Types: 419
- HP Printers: 1173
- Apache : 1104
- IIS: 529

Web Server Types

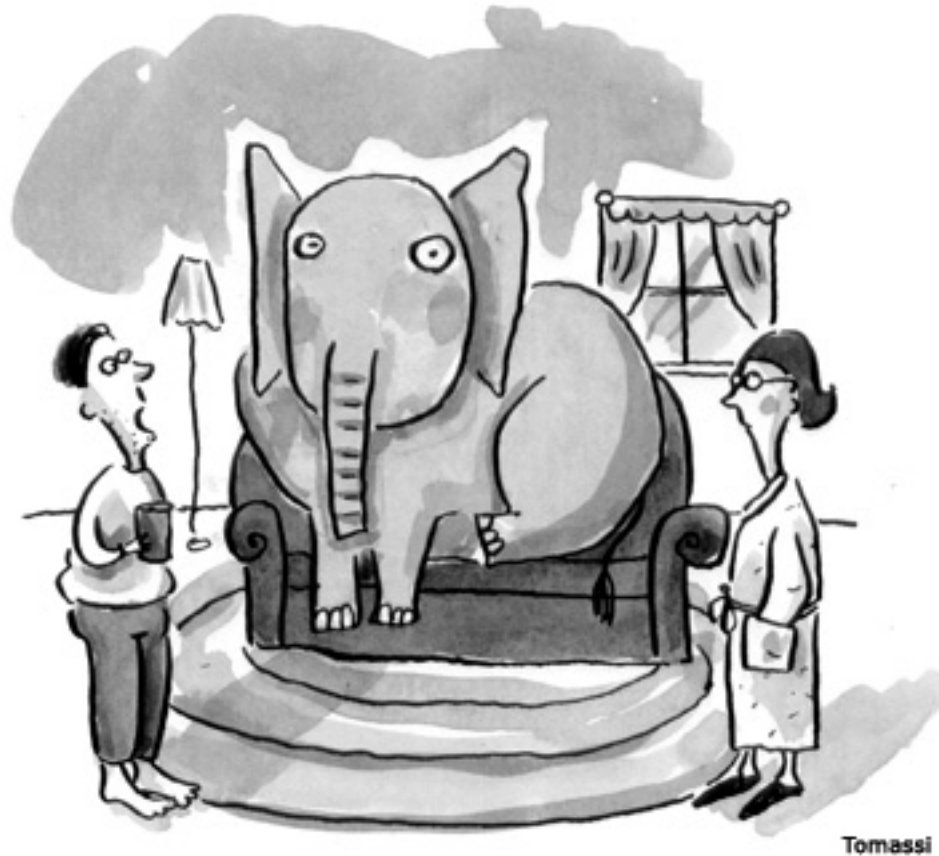


160 Server: Agranat-EmWeb/R5_2_6
194 Server: Microsoft-IIS/5.0
209 Server: Apache
217 Server: GoAhead-Webs
246 Server: Microsoft-IIS/6.0
519 Server:HTTP/1.0
902 Server: Virata-EmWeb/R6_2_1

Summary

- robots.txt – keep search engines from walking your web server
- **DOES NOT PROTECT DATA**
- Can divulge vulnerable / ignored files
- “Quiet Please” sign @ library -> defines polite behavior, does not enforce

Questions?



"How long has THAT been there?"

<http://tinyurl.com/ymeyqg>

UNIVERSITY OF MINNESOTA

Updates

- Add search engine authentication
- Examples modified

- References:

<http://www.robotstxt.org/wc/exclusion.html>

<http://johnny.ihackstuff.com/index.php?module=prodreviews&func=showcontent&id=96>