Topic: Bot Attacks

Question by: Tom Wrosch

Jurisdiction: Oregon

Date: December 17, 2013

Jurisdiction	Question(s)
	We were wondering if anyone else is experiencing this problem or, even better, have solved this problem and would share their solution with us: We are seeing a major increase in "screen scraping" by bots of our business registry database. While some of these are legitimate (commercial and LEO) applications, a lot of the increase is linked to folks who are using the info to test stolen credit cards and use the info in other criminal activities. The bot searches are getting so numerous they are amounting to Denial of Service attacks, and have caused our system response times to slow to unacceptable levels at times. They say 61% of internet traffic is now attributable to bot searches. We are loath to lock down our systems so that no one can do searches without an account, but are finding it difficult to find good solutions. Anybody else have this problem and come up with a good solution?
Manitoba	
Corporations Canada	
Alabama	
Alaska	
Arizona	
Arkansas	
California	
Colorado	See additional comments below
Connecticut	
Delaware	
District of Columbia	We have not experienced this issue in the District of Columbia. The reason being is that we require account registration before any services can be accessed including data search.
Florida	

Jurisdiction	Question(s)
	We were wondering if anyone else is experiencing this problem or, even better, have solved this problem and would share their solution with us: We are seeing a major increase in "screen scraping" by bots of our business registry database. While some of these are legitimate (commercial and LEO) applications, a lot of the increase is linked to folks who are using the info to test stolen credit cards and use the info in other criminal activities. The bot searches are getting so numerous they are amounting to Denial of Service attacks, and have caused our system response times to slow to unacceptable levels at times. They say 61% of internet traffic is now attributable to bot searches. We are loath to lock down our systems so that no one can do searches without an account, but are finding it difficult to find good solutions. Anybody else have this problem and come up with a good solution?
Georgia	
Hawaii	
Idaho	
Illinois	
Indiana	
Iowa	
Kansas	Kansas has not experienced this problem.
Kentucky	
Louisiana	When we first went live with our online application we had a similar problem. We realized that we were using a query string parameter that someone could loop through that allowed them to scrape the public information displayed. We encrypted the parameter so that they couldn't do that.
Maine	We have been experiencing the same issues here in Maine. Our online partner Informe has been blocking IP addresses when they notice it, but it doesn't resolve the issue. It has been causing issues with our database on the front end and back end.
Maryland	
Massachusetts	
Michigan	I spoke with our IT Staff/Security Team and they indicated that they have had no incidents reported to them regarding screen scrapping. However, the assumption is that if people want our data (and it's not id/password protected) then our screens are probably being scraped. The security team did pose some solutions to this, such as implementing user ids and passwords or adding a captcha to the application screen. I do recall a couple of years ago one report of screen scraping in Michigan but have not heard anything since.

Jurisdiction	**Question(s)**
	We were wondering if anyone else is experiencing this problem or, even better, have solved this problem and would share their solution with us: We are seeing a major increase in "screen scraping" by bots of our business registry database. While some of these are legitimate (commercial and LEO) applications, a lot of the increase is linked to folks who are using the info to test stolen credit cards and use the info in other criminal activities. The bot searches are getting so numerous they are amounting to Denial of Service attacks, and have caused our system response times to slow to unacceptable levels at times. They say 61% of internet traffic is now attributable to bot searches. We are loath to lock down our systems so that no one can do searches without an account, but are finding it difficult to find good solutions. Anybody else have this problem and come up with a good solution?
Minnesota	
Mississippi	
Missouri	
Montana	
Nebraska	
Nevada	See additional comments below
New Hampshire	
New Jersey	
New Mexico	
New York	

Jurisdiction	Question(s)
	We were wondering if anyone else is experiencing this problem or, even better, have solved this problem and would share their solution with us: We are seeing a major increase in "screen scraping" by bots of our business registry database. While some of these are legitimate (commercial and LEO) applications, a lot of the increase is linked to folks who are using the info to test stolen credit cards and use the info in other criminal activities. The bot searches are getting so numerous they are amounting to Denial of Service attacks, and have caused our system response times to slow to unacceptable levels at times. They say 61% of internet traffic is now attributable to bot searches. We are loath to lock down our systems so that no one can do searches without an account, but are finding it difficult to find good solutions. Anybody else have this problem and come up with a good solution?
North Carolina	I asked our IT staff and they said it had gotten to the point where 30% of our web traffic was due to bots mostly (google, bing, etc.). They went through the log and figured out which pages were being hit repeatedly and added them to the robots.txt file at the base of the web site. ex (robots.txt): # www.secretary.stat.nc.us robots.txt User-agent: * Disallow: /boardnotices/board.aspx Disallow: /BoardNotices/board.aspx Disallow: /BoardNotices/Board.aspx Disallow: /corporations/searchresults.aspx ... After that we only got 1 complaint from a legitimate screen scraper and we were able to change robots.txt to accommodate him. You can also specify specific user agents. Our IT staff doesn't think this would help with the credit card stealing situation, since those users are probably not appearing as bots, but it should help the web traffic.
North Dakota	
Ohio	
Oklahoma	
Oregon	
Pennsylvania	
Rhode Island	

Jurisdiction	Question(s)
	We were wondering if anyone else is experiencing this problem or, even better, have solved this problem and would share their solution with us: We are seeing a major increase in "screen scraping" by bots of our business registry database. While some of these are legitimate (commercial and LEO) applications, a lot of the increase is linked to folks who are using the info to test stolen credit cards and use the info in other criminal activities. The bot searches are getting so numerous they are amounting to Denial of Service attacks, and have caused our system response times to slow to unacceptable levels at times. They say 61% of internet traffic is now attributable to bot searches. We are loath to lock down our systems so that no one can do searches without an account, but are finding it difficult to find good solutions. Anybody else have this problem and come up with a good solution?
South Carolina	
South Dakota	
Tennessee	
Texas	
Utah	
Vermont	
Virginia	
Washington	
West Virginia	
Wisconsin	
Wyoming	

Additional comments:

Thomas Ose, Ose Micro Solutions, Inc. wrote:

Another alternative is to provide the data that clients want in a non-web format (such as XML or JSON) so that legitimate clients can access the information. Then you can control the crawlers and possibly have legitimate clients register or sign up for the alternate method. You can then also lock down your sites to all web crawlers that do not play nicely without effecting your users. A side benefit is that you would see no impact on performances on your sites and no denial of service because of someone hammering your web site.

COLORADO:

From our experience, there is the “people that play nice” way, and the “impolite people that ignore us” way.

The first is pretty easy, the second requires more attention (and attention probably means resources).

First off, do your website policies (privacy, disclaimers, legalese, etc.) tell people that data mining is not allowed and you reserve the right to take action? Here’s that section of our website Ts and Cs:

This website is intended for use by natural persons in obtaining information provided by the Secretary of State. Use of computerized “robots” or “data mining” of the information and images presented here is prohibited. Misuse of this website is prohibited and may result in the revocation of access to those persons or organizations using this site in a way not intended by the Secretary of State.

For people that play nice, your web folks can follow the conventions for good web crawlers. See http://www.robotstxt.org/ for pretty detailed information on helping well-behaved robots understand and play by the rules. All the legitimate web crawlers (google, bing, etc.) respect those directions. Your folks are probably already doing this.

For people that don’t play nice, you have to be able to identify them. Here are a couple techniques we use:

- Include a reference in your robots.txt file to a dummy web page (i.e., a page that is not referenced anywhere else on your website) that you say should not be indexed. Then keep track of the IP addresses that visit that dummy page. Block them with a firewall rule. The thinking is, the only robots that would visit that dummy page are the ones that checked the robots.txt and then ignored the guidance!

- Automatically or manually identify high-use IPs and subnets that access your site. One way of tracking these down is to look for computers with a low number of visits and a high number of page views. They don’t come very often, but when they do come, they hit a lot of web content. Block them with a firewall rule. You can do this manually, or Intrusion Detection Systems/Intrusion Prevention Systems (IDS/IPS) can be configured to do it automatically.

- Hey, now you get phone calls from legitimate customers that aren’t scraping, but have programmed filing and searching apps to conduct legit business! That’s okay, unblock them. Have your network/firewall folks make note of the IPs and subnets those legitimate customer are coming from, so you don’t block them tomorrow or next week when you next review traffic for “bad actors”. IDS/IPS systems that put blocks in place automatically can also handle a list of exclusions from automatic action.

- When a legit customer changes their Internet Service Provider, you’ll probably quickly start blocking them because their IP addresses changed from the one you knew about and they are pushing high volume. We know from experience that they will call us pretty quickly when they can’t get to our website, and we will add the new IP range to our exclusions list.

- When we block, we keep track of when we put the block in place. After a month, we unblock that IP or subnet. If they start hammering us again, we'll put the block back in place. Maybe forever. Or at least until we go back and do a comprehensive firewall rule review, about once a year.

- If you've got the money, invest in a web application firewall. These devices learn what "normal" web traffic looks like. When they see stuff that does not look normal, they automatically block traffic.

NEVADA:

In Nevada, we had issues where screen scraping and automated services were slowing our systems significantly. We have taken steps to ensure that our systems are not compromised or brought to their knees by bot or other automated services.

We have several known uses of "bots"

1. We have bots that use the free business search to pull up info and then save the information from the screen.

a. We use the equivalent of a Data Warehousing to handle the load. Requests from the free business search are routed to a specific server to handle the request while searches within our internal systems are sent to another. We replicate the data from the primary eSoS server to the "warehouse" so the information might be slightly out of sync when we are having technical issues. The short version is that anonymous searches are sent to one server and searches from people logged into our Portal go to another. This has worked well for many years.

2. We have bots that go through all our online forms daily and download them to make sure they have the latest forms. We have a disclaimer stating that use of automated process is not allowed but I doubt anyone pays attention to that.

3. We have bots that we allow to access our legacy site to automate the RA business processes for annual list and articles filings. The bot come from specific IP addresses that have been allowed to access (whitelisted) our legacy site. This is on a very limited basis and agreements are in place for those authorized users. We are currently developing portal services to allow these automated services and then we will retire our legacy system. IP Addresses that do negatively affect our systems are identified and blacklisted.

In addition to the methods we use to control security threats, we also take advantage of EITS' (State information Technology Department) security capabilities. We use the State's BIG IP F5 switches for load balancing and traffic monitoring. We have strictly adhered to the State OIS's recommendations on segmentation in our architecture which includes multi-tiered application, security zones, and hardware deployment. From security measures implemented by the State, all the way down to our development, we have taken every step feasible to secure our system and isolate potentially dangerous traffic. We never feel completely secure but we make security part of everything we do. We also use Captcha which is very helpful in keeping bots out of the system. Customers will complain but at the same time the systems will be available when the need them. I'm trying to keep this as high level as possible because it is a very complex architecture developed over time.

Full text of email:

We were wondering if anyone else is experiencing this problem or, even better, have solved this problem and would share their solution with us:

We are seeing a major increase in "screen scraping" by bots of our business registry database. While some of these are legitimate (commercial and LEO) applications, a lot of the increase is linked to folks who are using the info to test stolen credit cards and use the info in other criminal activities. The bot searches are getting so numerous they are amounting to Denial of Service attacks, and have caused our system response times to slow to unacceptable levels at times.

They say 61% of internet traffic is now attributable to bot searches. We are loath to lock down our systems so that no one can do searches without an account, but are finding it difficult to find good solutions. Anybody else have this problem and come up with a good solution?

Thanks,

Tom

Tom Wrosch
Registry Programs Administrator
Oregon Secretary of State Corporation Division
255 Capitol St. NE Ste. 151
Salem, OR 97310

(503) 986-0511