Botbusters Handbook
Peter Katz
VP, Customer Lifecycle Solutions Support, Elsevier
What’s the problem and how (can) we fix it?
Is there a problem with terminology

“BOT Usage”
- Sounds innocuous... it isn’t
- BOT usage is a term used to cover all types of systematic mining of our content
- Some BOTs are TDM workers
- Most BOTs are thieves

“IP Authentication”
- is a misnomer
- We use IP* access as a mutually convenient method of delivering our product and content to our customers
- No authentication is involved

Bots are exploiting IP access and our lack of authentication to steal our content

*Our customers often use a proxy server, such as OCLC’s EZProxy, for managing patron access to our content
OLD NEWS?

SCI-HUB

...to remove all barriers in the way of science

enter URL, PMID / DOI or search string
Stealing and storing stolen content

* Data from Elsevier ISDP analysis, Alexa public web analytics and sci-hub.io

- 500+ sites link to it
- Contains 62M+ papers
- Iran, Russia, China, Brazil, and Egypt are top 5 visitors
- Identified 200 unique universities that had proxy accounts used on behalf of Sci-Hub
- *Science* study indicated that 88% of those surveyed think it isn’t wrong to download pirated papers *(Who's downloading pirated papers? Everyone)*

- Levine-Clark, McDonald & Price determined that 260 of 300 articles (100@ Social Sci, Arts & Humanities, Life Sciences) available in Sci-Hub. *(https://www.slideshare.net/pricein/access-to-freely-available-journal-articles-gold-green-and-rogue-open-access-across-the-disciplines)*
Are your students & faculty’s credentials being stolen for sale and re-use?
Welcome to "Scientific Journals For All" - Digital Library

This page has been created in order to share access to databases with thousands of scientific journals and databases that is often difficult to access for many university students and health professionals of different countries around the world. Enjoy it!

  *Don't forget to use Google Chrome browser for all of our resources.

About Us:

- **We are a group dedicated to share access to scientific information**
  - Access to hundreds of databases and thousands of scientific journals.

- **Periodical updates and publications about scientific access**
  - We have a facebook page where you will find our resources.

- **Scientific articles to everyone for FREE**
  - We have a facebook group where you can request your scientific articles-look in the tabs above.

- **Access Codes**
  - We have a access codes system to access to different resources.

Copyright © 2016 Scientific Journals For All - Digital Library
And there's more!

http://3g0317.com/
http://yanshi9.com/
retrieved 26 Oct 2015
Iranian sites offering publisher content
روش ششم جستجو و دانلود از طریق خود پایگاه اطلاعاتی

برای هر پایگاه اصلی سرویسی طراحی شده است که کاربر با وارد شدن به آن گویی در پایگاه اصلی می‌باشد، با این تفاوت که زیر مجموعه‌ای از دانش‌گستری می‌باشد و کاربر می‌تواند با یک کلیک بر روی آیکون دانلود اقدام به تهیه مقاله مورد نیاز خود کند. برای مثال برای پایگاه Elsevier به آدرس زیر بروید:

- http://sciedirect.daneshgostar.org

قابلیت این سرویس این است که کاربر می‌تواند از امکانات جستجو بی‌پرهفته هر پایگاه اطلاعاتی استفاده نماید و خیلی سریع مقاله خود را دانلود کند. علاوه بر IEEE و Wiley، Springer با استفاده از لینک‌های زیر به راحتی می‌توانید از پایگاه‌های استفاده کنید:

- http://springer.daneshgostar.org
- http://wiley.daneshgostar.org
- http://IEEE.daneshgostar.org
امروز، دنیای علم در شاهد است. از سوئدیونی که به بینهایت اعضا کرده ساده و پیش‌ساخته برابر شما در راه‌های میزانی که گزارش دارد. این را باید به‌دینا که اگر به‌دینا ودیع دارد، اصولی است که تویت‌های به این دستیابید، اگر دست‌نبردیدی، به ایاد و راه‌های به‌دینا عصب شده است. ردیابی‌ها و راه‌های به‌دینا که شما در راه یک روسی به‌دینا که فهمیده‌اند، من راه را به شما نشان می‌دهم. پیام‌های خوشنام است.

از دیقیقه به دیقیقه امنیت متشکل رودم، اما با خودم می‌گفتم چرا نزن، اینک خسته‌ای رو نمی‌ده، نابه‌پک یکی روش مثل یکی قصر، قلی از کاتی منتسب به کلیه. محمد بنر توجه، طالب داراد بر جا عزیز در ویلاگ ممکن است دانشجوی سراسری شود.
How does it happen?

Sai Soe <sai.soe@sydney.edu.au>
mer. 2016-01-06 19:47

UNIVERSITE LAVAL

We have just observed an unusual Sign-In to your mail

Click Here To Confirm Your Login

ICT Help Desk
After chasing the bots for the past two years, I have the following observations

• As Rick Anderson has written, compromised/stolen/shared credentials affects much more than the ability to download content.
• Most of the intrusions originate in China, India, Iran and former Soviet states. Sci-hub copycats are being built in all these regions; many charge for access.
• 90% of the incidents are generated by unaffiliated users who have gained access to university's network for the purpose of stealing content
• EndNote can be used to crawl content
• We typically see this activity before the customer does: while EZproxy has a mechanism for flagging extraordinarily high usage or traffic that is malicious it is difficult for customers to configure since each provider may have different rules.
• Most customers are responsive when they are notified of the issue but it can take several days (with attendant loss of content) before they can locate the source and block the username.
• Some customers do not have good IT support and are not able to locate the source of crawling: some don’t have active logging on their systems; some have claimed there can be a privacy issue
• Once a username is blocked, others are used, therefore is often very difficult to stop malicious activity
• Some customers seem particularly prone to attack: they are known for having large collections which are more attractive to thieves
What is this bot chasers routine?

- Review daily reports

<table>
<thead>
<tr>
<th>Alert Start Time</th>
<th>Alert End Time</th>
<th>In Ip Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>5/31/2016 2:02:04 AM</td>
<td>5/31/2016 2:38:46 AM</td>
<td>False 488</td>
</tr>
<tr>
<td>5/31/2016 2:36:02 AM</td>
<td>5/31/2016 2:39:24 AM</td>
<td>False 75</td>
</tr>
<tr>
<td>5/31/2016 8:15:46 PM</td>
<td>5/31/2016 8:24:30 PM</td>
<td>False 182</td>
</tr>
<tr>
<td>5/31/2016 0:04:01 PM</td>
<td>5/31/2016 8:11:19 PM</td>
<td>False 171</td>
</tr>
<tr>
<td>5/31/2016 9:27:12 AM</td>
<td>5/31/2016 10:00:52 AM</td>
<td>True 1,214</td>
</tr>
<tr>
<td>5/31/2016 10:01:00 AM</td>
<td>5/31/2016 12:13:47 PM</td>
<td>True 1</td>
</tr>
</tbody>
</table>

- Take a deeper dive

<table>
<thead>
<tr>
<th>Event Request Date</th>
<th>Event Id &amp; Account Number</th>
<th>Account Name</th>
<th>Ip Address</th>
<th>Session Id</th>
<th>Article Form</th>
<th>Content Cols</th>
<th>Content Cols</th>
<th>Journal Title</th>
<th>Article Title</th>
<th>Request Start Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>5/31/16</td>
<td>004258-272.html</td>
<td>serial</td>
<td>JL</td>
<td>12:06:01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5/31/16</td>
<td>01234590-007.html</td>
<td>serial</td>
<td>JL</td>
<td>12:00:35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5/31/16</td>
<td>03235690-007.html</td>
<td>serial</td>
<td>JL</td>
<td>12:00:35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5/31/16</td>
<td>05678900-007.html</td>
<td>serial</td>
<td>JL</td>
<td>12:00:35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5/31/16</td>
<td>00567890-007.html</td>
<td>serial</td>
<td>JL</td>
<td>12:00:35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5/31/16</td>
<td>07567890-007.html</td>
<td>serial</td>
<td>JL</td>
<td>12:00:35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Reach out to customer

Dear Ms.

Our usage tracking system has recently alerted us to abnormally high usage of Science Direct at xxxx the last two days.

These activities show a series of systematic querying and subsequent downloading of articles from SD. While querying

According to our terms & conditions and supported as a general function of Science Direct, the systematic
downloading.

Most of the activity came via one IP address and was focused on journals.
Where do we go from here

• Determine how the industry can work more effectively together to reduce content theft
• Education! FAQs for improved security as well as how to find intruders
• Build a better alliance with customers and OCLC to improve proxy security
• New access methods