2018

Web Honeypots for Spies

Henderson, Blake T.; McKenna, Sean F.; Rowe, Neil C.

Web Honeypots for Spies

Blake T. Henderson, Sean F. McKenna, and Neil C. Rowe (presenter)

U.S. Naval Postgraduate School
Monterey, California, USA
ncrowe@nps.edu

Motivation

- We are building honeypots for document-collecting spies who are searching the Web for intelligence information.
- It is important for governments, organizations, and businesses to know who is accessing their public documents.
- Further, we may be able to assess the relative degree of interest elicited by users in documents.
- One experiment of ours set up a site with bait documents and used two site-monitoring tools, Google Analytics and AWStats, to analyze the traffic.
- Another experiment of ours analyzed bot traffic on a similar real site, the library site at our school.

Previous honeypot research

- Honeypots have been used from the early days of cybersecurity.
- We have run honeypots for many years at our school on lines outside the School firewall.
- They are a good way to collect cyberattack intelligence.
- However, they need to be shaped because different attackers are interested in different things.
- We have run SSH honeypots, Web honeypots, industrial-control system honeypots, and several other kinds.

The honeypot we set up

NAVAL POSTGRADUATE SCHOOL

Dudley Knox Library

Naval Postgraduate School Future Research

The Naval Postgraduate School Future Research Department is dedicated to exploring and funding graduate level research for the DoD's top priorities for national defense.

Air

A compilation of requisite expanded research requirements to maintain air superiority.

Budget

A compilation of budget requirements and restraints through 2027.

Cyber

Advanced cyber research and associated source code in preparation for cyberwar planning.

Declassified Projects

A compilation of formerly classified case studies.

Policy

Foreign policy challenges with an emphasis on China and Russia.

Science & Technology

A compilation of scientific and technological research opportunities to ensure a persistent competitive advantage.

Space

An anthology of state-of-the-art spacecraft research for enduring global strike and intelligence capabilities.

Special Operations

A compilation of crucial technical and operational requirements needed in order to protect and advance our Nation's interests.

Subsurface

A collection of research on manned and unmanned submarine technologies.

Surface

Explorations into advancing surface ship design and global reach capabilities.

Weapons Systems

An investigation into leveraging cutting-edge technology for weapons systems in the 21st Century.

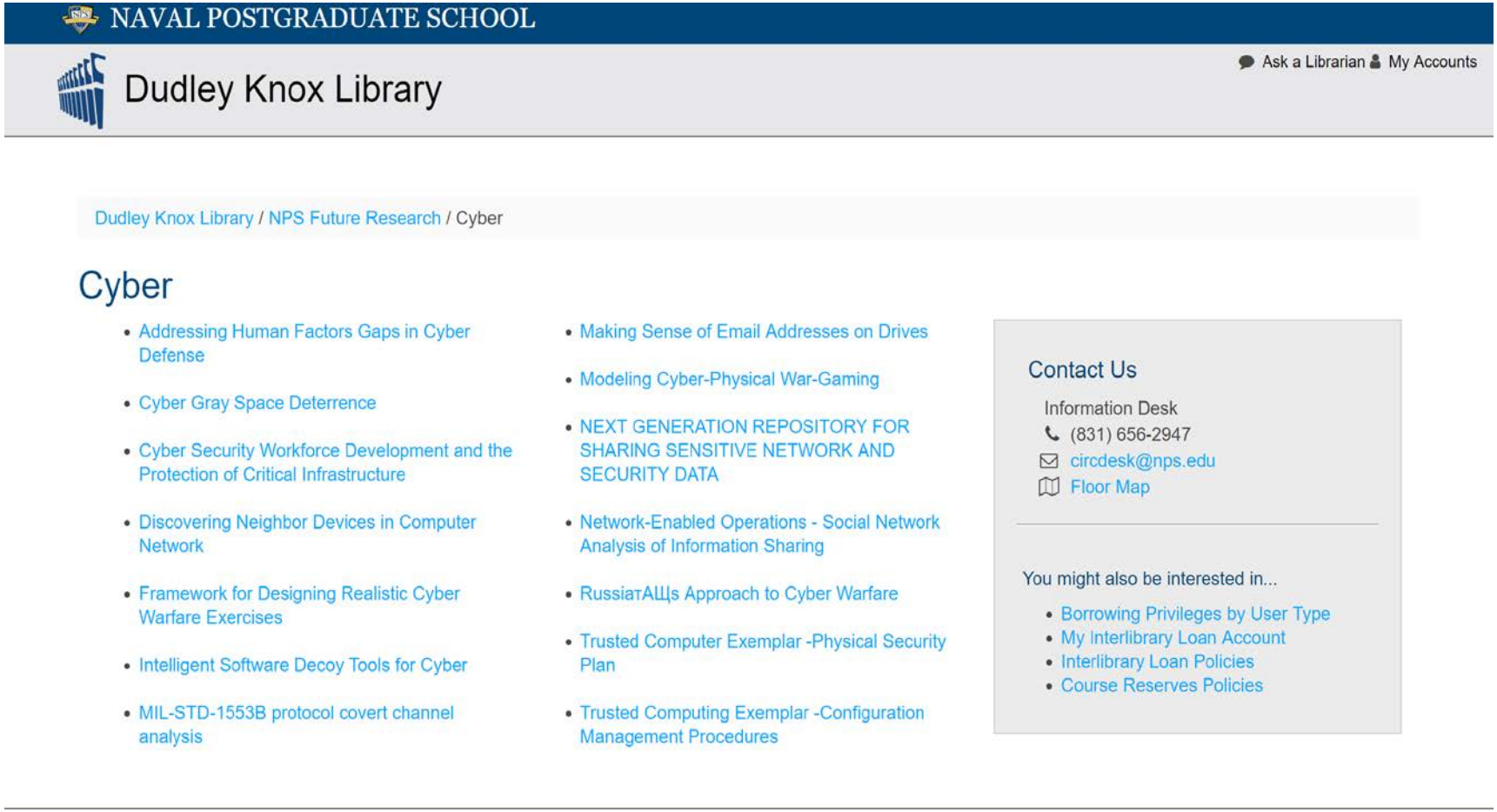
Contact Us

Information Desk
(831) 656-2947
circdesk@nps.edu
Floor Map

You might also be interested in...

- Borrowing Privileges by User Type
- My Interlibrary Loan Account
- Interlibrary Loan Policies
- Course Reserves Policies

Example subpage



NAVAL POSTGRADUATE SCHOOL

Dudley Knox Library

Ask a Librarian My Accounts

Dudley Knox Library / NPS Future Research / Cyber

Cyber

- Addressing Human Factors Gaps in Cyber Defense
- Cyber Gray Space Deterrence
- Cyber Security Workforce Development and the Protection of Critical Infrastructure
- Discovering Neighbor Devices in Computer Network
- Framework for Designing Realistic Cyber Warfare Exercises
- Intelligent Software Decoy Tools for Cyber
- MIL-STD-1553B protocol covert channel analysis
- Making Sense of Email Addresses on Drives
- Modeling Cyber-Physical War-Gaming
- NEXT GENERATION REPOSITORY FOR SHARING SENSITIVE NETWORK AND SECURITY DATA
- Network-Enabled Operations - Social Network Analysis of Information Sharing
- RussiaтАЩs Approach to Cyber Warfare
- Trusted Computer Exemplar -Physical Security Plan
- Trusted Computing Exemplar -Configuration Management Procedures

Contact Us

Information Desk
(831) 656-2947
circdesk@nps.edu
Floor Map

You might also be interested in...

- Borrowing Privileges by User Type
- My Interlibrary Loan Account
- Interlibrary Loan Policies
- Course Reserves Policies

Design of the document honeypot

- We set up a Web server on what appeared to be a School address and monitored its traffic.
- We could not use a real School address, but used one listed as being owned by the School.
- We also used graphics and layout typical of the School library.
- We selected 132 unclassified documents in currently popular fields of interest covered by the U.S. Department of Defense.
- Documents were 11 areas; most published 5-10 years ago.
- Our server: Ubuntu Linux and Apache 2.4.18 on Dell workstation hardware.
- Ports: 80 for web traffic and 22 for SSH.
- We registered our domain name with Google to get it indexed.

Usage monitoring software

- Google Analytics
  - Counts site page visits, time on a page, general geographic information about visitor IP.
  - Requires a tracking ID on the honeypot home page.
  - We also created an event trigger to record downloads of documents.
  - Tries to exclude bot traffic from statistics. Bots can be legitimate like Google's indexing, but some are malicious.
- AWStats
  - Measures similar things as Google Analytics.
  - Does not exclude bots, which gave much more data.
  - Not as sophisticated at Google Analytics in providing breakdowns of visitors by document.

General honeypot statistics

- We ran for 5.5 months.
- The home page had 91.1% of the page views according to AWStats.
- There were 87 attempts to use our site as a proxy, mostly to Chinese sites.

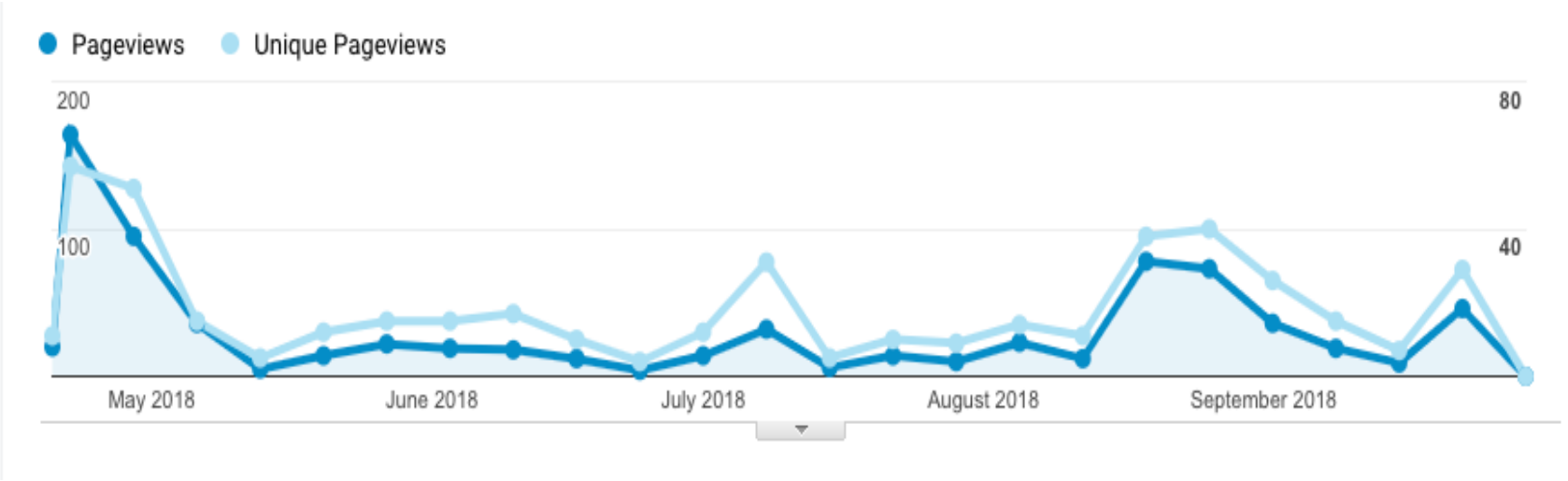
Most popular pages by Google Analytics

	Event Label	Total Events ↓	Total Events
		72 % of Total: 100.00% (72)	72 % of Total: 100.00% (72)
1.	Advanced Aerobots for Scientific Exploration	4	5.56%
2.	Bioeffects on an In Vitro Model by Small-Scale Explosives and Shock Wave Overpressure Impacts	3	4.17%
3.	CIA Sculpture Study Group	3	4.17%
4.	Effectiveness of the CIA Counterterrorist Interrogation Techniques	3	4.17%
5.	A HYBRID AGENT APPROACH FOR SET-BASED CONCEPTUAL SHIP DESIGN	2	2.78%
6.	Applied Explosives Technology	2	2.78%
7.	Discovering Neighbor Devices in Computer Network	2	2.78%
8.	F-35 Alternate Engine Program	2	2.78%
9.	F-35 Aluminum Composite Stack Drilling	2	2.78%
10.	High energy solid state and free electron laser systems in tactical aviation	2	2.78%
11.	John Nash Letters	2	2.78%
12.	MIL-STD-1553B protocol covert channel analysis	2	2.78%
13.	Multimodal Displays in Army Human-Robot Operations	2	2.78%
14.	Safe Haven Configurations for Deep Space Transit Habitats	2	2.78%

Most popular pages by AWStats

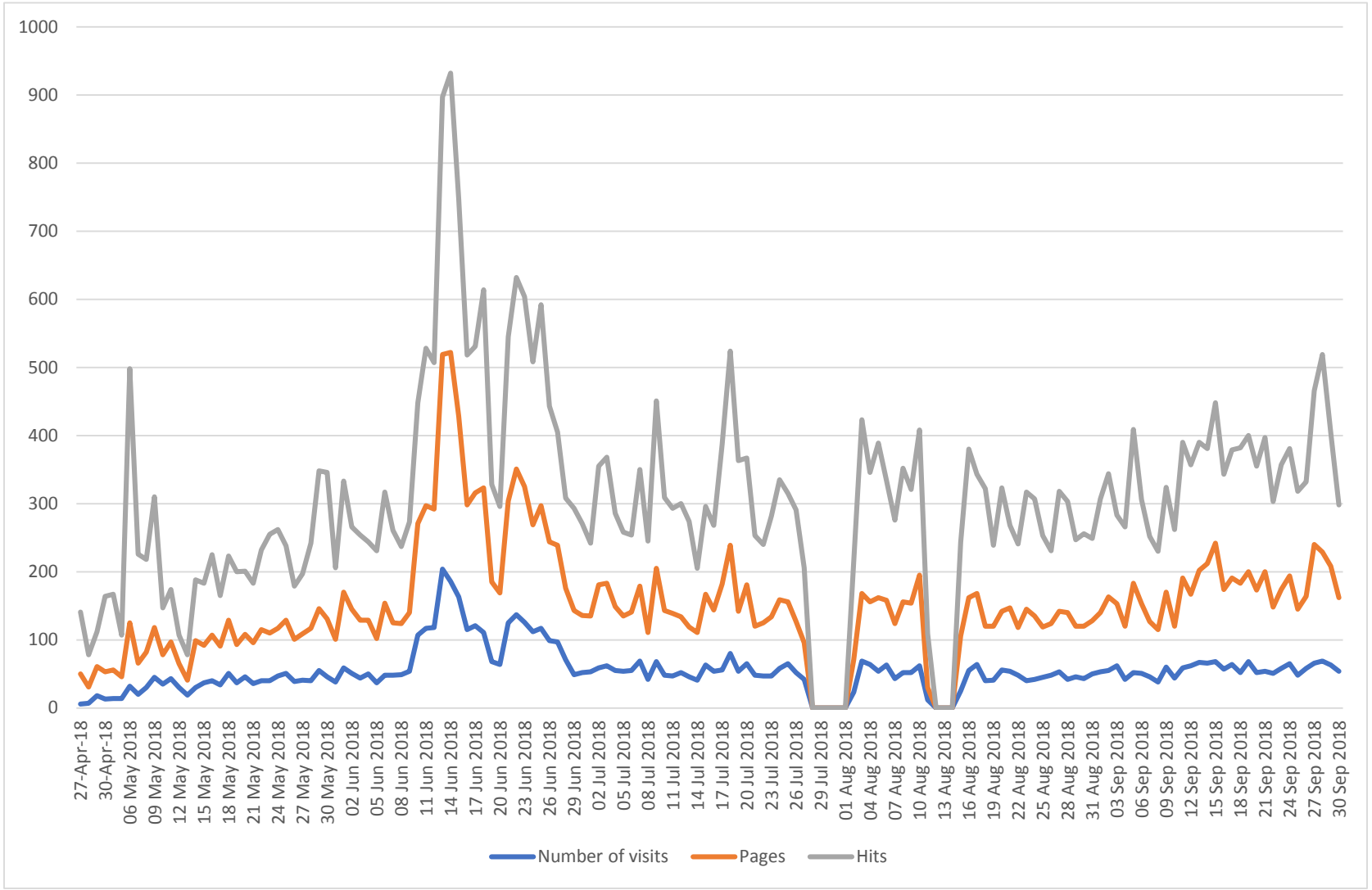
Category	Top 10 Visits	Sum of Hits	Sum of In-complete Hits
Science & Technology	Multi-Task Convolutional Neural Network for Pose-Invariant Face Recognition	591	328
Surface	Hydrostatic and hydrodynamic analysis of a lengthened DDG-51	207	104
Surface	DDG-1000 missile integration	182	211
Policy	China's evolving foreign policy in Africa	149	10
Surface	Establishing the Fundamentals of a Surface Ship Survivability Design Discipline	130	220
Special Operations	Roles of Perseverance, Cognitive Ability, and Physical Fitness - U.S. Army Special Forces	128	19
Surface	A Salvo Model of Warships in Missile Combat Used to Evaluate Staying Power	110	411
Cyber	MIL-STD-1553B protocol covert channel analysis	109	72
Policy	Analysis of government policies to support sustainable domestic defense industries	92	16
Policy	Russia's natural gas policy toward Northeast Asia	89	421

Activity over time according to Google Analytics

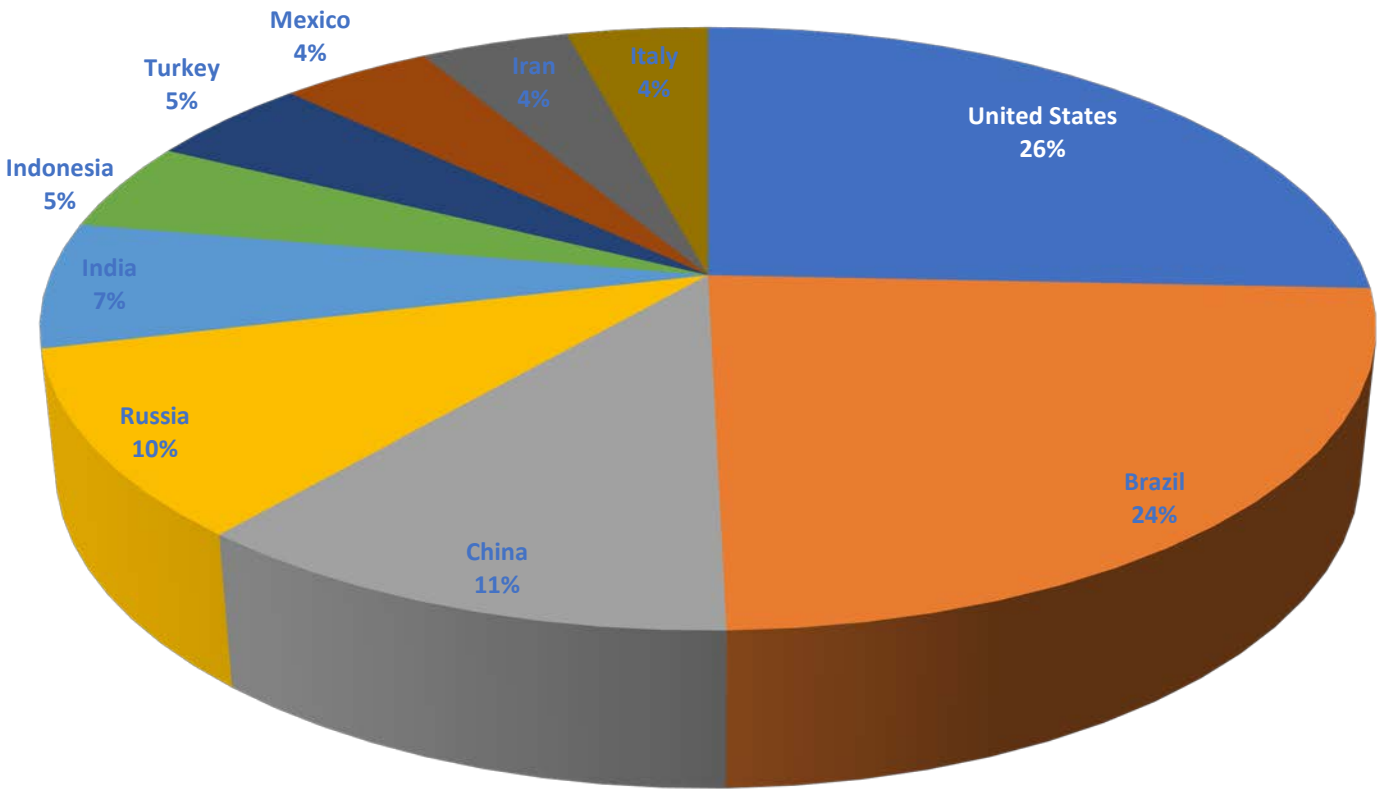


- The initial burst is typical of new honeypots.
- Other swells likely represent “campaigns” or organized querying.
- The September peak is due to a questionnaire we administered to human subjects about our site.
- AWStats had a peak more towards June not present above, for bot campaigns.

Activity over time according to AWStats



Breakdown of users by country



Analytics on our real School library site

- The previous work indicated most users were bots. So it is useful to analyze the activities of bots alone.
- We wrote a “sandtrap” script to capture bot resource requests at our library.
- This was implemented as a server-side PHP script because our site uses PHP.
- We logged the time, IP address, and user agent of the visitors for five weeks.
- The library was particularly interested in bots looking for email addresses, so we created some pages with link text containing addresses.
- We also set up a robots.txt file to request avoidance of certain pages, and checked whether bots respected that.

Preliminary experiments with sample crawlers

Program	Robots.txt	Allowed Resource (class)			Banned Resource (noclass)		
	Check-ed?	.pdf	.doc	.html	.pdf	.doc	.html
Import.io	No	No	Yes	Yes	No	Yes	Yes
80Legs	Yes	No	No	Yes	No	No	No
Scrapy	No	Yes	Yes	Yes	Yes	Yes	Yes
Selenium	No	No	No	No	No	No	No
ScrapeBox	No	No	Yes	Yes	No	Yes	Yes
iRobotSoft	No	No	No	Yes	No	No	Yes
Anenome	No	No	Yes	Yes	No	Yes	Yes

Selenium was the best, but it does not scale well. The others were not very respectful of robots.txt.

Overall statistics on Web logs

	Human Traffic	Bot Traffic
Total Requests	334,673	596,028
Average Req/Day	9843	17530
Bandwidth Consumed	179.74 GB	39.46 GB
% of Distinct Requests	35.96 (36%)	64.04 (64%)

We distinguished human from bot traffic by extracting the “User-Agent” field of HTTP headers and comparing it to Splunk’s keyword list of bot names.
However, this field is easy to spoof and won’t identify malicious bots.

More statistics

- 46 self-identifying bots visited the site using 505 different addresses.
- Google, Yahoo, and Bing accounted for 99% of the search requests.
- Of 358 requests for files, 216 were for the unrestricted folder, 142 were for the restricted class folder.
- Unrestricted folder: We observed 21 Web bot campaigns from 59 IP addresses with 216 resource requests. 11 of these (52%) used forged user-agent strings.
- Restricted folder: 16 Web bot campaigns from 25 IP addresses with 142 resource requests. 7 used forged user-agent fields.
- 40 IPs were in Project Honeypot's blacklist, but none of these requested resources.

Conclusions

- Intelligence gathering is facilitated by the World Wide Web.
- It also appears easy to fool intelligence gathering with honeypots.
- We have shown that it suffices to monitor this activity with a few simple tools.
- Bot activity is scattered over topics, suggesting that most retrievals are done by relatively indiscriminate bots that conceal the real interests of human users. Thus, attempts to offer bait were ineffective.
- However, some keywords like “neural”, “DDG”, and “China” attracted a bit more traffic.
- Results also showed that content-specific anchors were useful in detecting bots, and that bots often did not often respect site terms of service.