Deep Web

Part II.B. Techniques and Tools:
Network Forensics

CSF: Forensics Cyber-Security
Fall 2015
Nuno Santos
Summary

- The Surface Web
- The Deep Web
Our journey in this course:

- Part I: Foundations of digital forensics
- Part II: Techniques and tools
  - A. Computer forensics
  - B. Network forensics
  - C. Forensic data analysis
Previously: Three key instruments in cybercrime

- **Anonymity systems**
  How criminals hide their IDs

- **Botnets**
  How to launch large scale attacks

- **Digital currency**
  How to make untraceable payments
Today: One last key instrument – The Web itself

- Web allows for accessing **services** for criminal activity
  - E.g., drug selling, weapon selling, etc.

- Provides huge **source of information**, used in:
  - Crime premeditation, privacy violations, identity theft, extortion, etc.

- To find services and info, there are powerful **search engines**
  - Google, Bing, Shodan, etc.
The Web: powerful also for crime investigation

- **Powerful investigation tool** about suspects
  - Find evidence in blogs, social networks, browsing activity, etc.

- **The playground** where the crime itself is carried out
  - Illegal transactions, cyber stalking, blackmail, fraud, etc.
The sophistication of offenses (and investigations) is driven by the **nature** and **complexity** of the Web.
The web is deep, very deep...

- What’s “visible” through typical search engines is minimal

Surface Web (4% of WWW content)

Deep Web (96% of WWW content)
What can be found in the Deep Web?

- **Deep Web** is not necessarily bad: it’s just that the content is not directly indexed.

- Part of the deep web where criminal activity is carried out is named the **Dark Web**.
Some examples of services in the Web “ocean”
Offenders operate at all layers

Investigators too!
The Surface Web

The Deep Web
The Surface Web
The **Surface Web** is that portion of the World Wide Web that is readily available to the general public and searchable with standard web search engines.

- AKA Visible Web, Clearnet, Indexed Web, Indexable Web or Lightnet

As of June 14, 2015, Google's index of the surface web contains about 14.5 billion pages.
Surface Web characteristics

- **Distributed data**
  - 80 million web sites (hostnames responding) in April 2006
  - 40 million active web sites (don’t redirect, …)

- **High volatility**
  - Servers come and go …

- **Large volume**
  - One study found 11.5 billion pages in January 2005 (at that time Google indexed 8 billion pages)
Surface Web characteristics

- **Unstructured data**
  - Lots of duplicated content (30% estimate)
  - Semantic duplication much higher

- **Quality of data**
  - No required editorial process
  - Many typos and misspellings (impacts IR)

- **Heterogeneous data**
  - Different media
  - Different languages
As of 2003, about 70% of Web content is images, HTML, PHP, and PDF files.
How to find content and services?

- Using search engines

1. A web crawler gathers a snapshot of the Web
2. The gathered pages are indexed for easy retrieval
3. User submits a search query
4. Search engine ranks pages that match the query and returns an ordered list
How a typical search engine works

- **Architecture of a typical search engine**

![Diagram of the architecture of a typical search engine](image)

- Users
- Interface
- Query Engine
- Crawler
- Indexer
- Web
- Index
- Lots and lots of computers
What a Web crawler does

- The Web crawler is a foundational species
  - Without crawlers, there would be nothing to search

- Creates and repopulates search engines data by navigating the web, fetching docs and files
What a Web crawler is

- In general, it’s a program for downloading web pages
  - Crawler AKA spider, bot, harvester

- Given an initial set of seed URLs, recursively download every page that is linked from pages in the set
  - A focused web crawler downloads only those pages whose content satisfies some criterion

- The next node to crawl is the URL frontier
  - Can include multiple pages from the same host
Crawling the Web: Start from the seed pages

Web

Seed pages

URLs crawled and parsed

URLs frontier

Unseen Web
Crawling the Web: Keep expanding URL frontier

- **Seed Pages**
- **URL frontier**
- **Unseen Web**
- **URLs crawled and parsed**
- **Crawling thread**
Web crawler algorithm is conceptually simple

- **Basic Algorithm**

  Initialize queue \( Q \) with initial set of known URL’s
  
  **Until** \( Q \) empty or page or time limit exhausted:
  
  - **Pop** URL, \( L \), from front of \( Q \)
  - **If** \( L \) is not to an HTML page (.gif, .jpeg, .ps, .pdf, .ppt…)
    - **continue** loop
  - **If** already visited \( L \), continue loop
  - Download page, \( P \), for \( L \)
  - **If cannot** download \( P \) (e.g. 404 error, robot excluded)
    - **continue** loop
  - Index \( P \) (e.g. add to inverted index or store cached copy)
  - Parse \( P \) to obtain list of new links \( N \)
  - Append \( N \) to the end of \( Q \)
But not so simple to build in practice

- **Performance**: How do you crawl 1,000,000,000 pages?
- **Politeness**: How do you avoid overloading servers?
- **Failures**: Broken links, time outs, spider traps.
- **Strategies**: How deep to go? Depth first or breadth first?
- **Implementations**: How do we store and update the URL list and other data structures needed?
Completeness
Is the algorithm guaranteed to find a solution when there is one?

Optimality
Is this solution optimal?

Time complexity
How long does it take?

Space complexity
How much memory does it require?
No single crawler can crawl the entire Web

- Crawling technique may depend on goal

- **Types** of crawling goals:
  - Create large broad index
  - Create a focused topic or domain-specific index
    - Target topic-relevant sites
    - Index preset terms
  - Create subset of content to model characteristics of the Web
    - Need to survey appropriately
    - Cannot use simple depth-first or breadth-first
  - Create up-to-date index
    - Use estimated change frequencies
Crawlers also be used for nefarious purposes

- Spiders can be used to collect email addresses for unsolicited communication
- From: http://spiders.must.die.net

**Spiders**

Crawling around the Web now are a wide variety of robots known as "spiders". Their goal is to recursively scan every document available and make a condensed form of the data they collected available to whoever controls them.

Many of these are Search Engines, which allow the general public to search the spider's index for specific items for "free". (Without the right tools you usually have to view their advertising, though.) On the whole, Search Engines are beneficial to have around.

However a new species of spider seems to have made its way onto the Web lately that is less beneficial to the public. These particular bugs have a more sinister purpose in mind: They collect e-mail addresses so that their owners can send everyone unsolicited advertising. No, thanks.
Open Source Crawlers in Java

**Heritrix**

Heritrix is the Internet Archive's open-source, extensible, web-scale, archival-quality web crawler project.

[Go To Heritrix](#)

**WebSPHINX**

WebSPHINX (Website-Specific Processors for HTML INformation eXtraction) is a Java class library and interactive development environment for Web crawlers that browse and process Web pages automatically.

[Go To WebSPHINX](#)

**JSpider**

A highly configurable and customizable Web Spider engine. Developed under the LGPL Open Source license, In 100% pure Java.

[Go To JSpider](#)

**WebEater**

A 100% pure Java program for web site retrieval and offline viewing.
A spider trap is a set of web pages that may be used to cause a web crawler to make an infinite number of requests or cause a poorly constructed crawler to crash. To “catch” spambots or similar that waste a website’s bandwidth.

Common techniques used are:

• Creation of indefinitely deep directory structures like
  • http://foo.com/bar/foo/bar/foo/bar/foo/bar/.....

• Dynamic pages like calendars that produce an infinite number of pages for a web crawler to follow

• Pages filled with many chars, crashing the lexical analyzer parsing the page
Search engines run specific and benign crawlers

- Search engines obtain their listings in two ways:
  - The search engines “crawl” or “spider” documents by following one hypertext link to
  - Authors may submit their own Web pages

- As a result, only static Web content can be found on public search engines

- Nevertheless, a lot of info can be retrieved by criminals and investigators, especially when using “hidden” features of the search engine
Google provides keywords for advanced searching

- Logic operators in search expressions
- Advanced query attributes: “login password filetype:pdf”

<table>
<thead>
<tr>
<th>Intitle, allintitle</th>
<th>Related</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inurl, allinurl</td>
<td>Phonebook</td>
</tr>
<tr>
<td>Filetype</td>
<td>Rphonebook</td>
</tr>
<tr>
<td>Allintext</td>
<td>Bphonebook</td>
</tr>
<tr>
<td>Site</td>
<td>Author</td>
</tr>
<tr>
<td>Link</td>
<td>Group</td>
</tr>
<tr>
<td>Inanchor</td>
<td>Msgid</td>
</tr>
<tr>
<td>Daterange</td>
<td>Insubject</td>
</tr>
<tr>
<td>Cache</td>
<td>Stocks</td>
</tr>
<tr>
<td>Info</td>
<td>Define</td>
</tr>
</tbody>
</table>
There’s entire books dedicated to Google hacking


*Ethical Hacking,*
http://www.nc-net.info/2006conf/Ethical_Hacking_Presentation_October_2006.ppt

A cheat sheet of Google search features:

A Cheat Sheet for Google Search Hacks -- how to find information fast and efficiently
A simple search: “cd ls .bash_history ssh”

Can return surprising results: this is the contents of a live .bash_history file
Google hacking examples: URL searches

- `inurl:` find the search term within the URL

Examples:
- `inurl:admin` to search for admin-related URLs
- `inurl:admin users mbox` to search for admin users' mailbox URLs
- `inurl:admin users passwords` to search for admin users' passwords URLs
filetype: narrow down search results to specific file type

filetype:xls "checking account" "credit card"
Google hacking examples: Finding servers

intitle:"Welcome to Windows 2000 Internet Services"

intitle:"Under construction" "does not currently have"
Google hacking examples: Finding webcams

- To find open unprotected Internet webcams that broadcast to the web, use the following query:
  - `inurl:/view.shtml`

- Can also search by manufacturer-specific URL patterns
  - `inurl:ViewerFrame?Mode=`
  - `inurl:ViewerFrame?Mode=Refresh`
  - `inurl:axis-cgi/jpg`
  - `...`
Google hacking examples: Finding webcams

- How to Find and View Millions of Free Live Web Cams

- How to Hack Security Cameras,

- How to Hack Security Cams all over the World
  [http://www.youtube.com/watch?v=9VRN8BS02Rk&feature=related](http://www.youtube.com/watch?v=9VRN8BS02Rk&feature=related)
And we’re just scratching the surface…

What can be found in the depths of the Web?
The Deep Web
Deep Web is the part of the Web which is not indexed by conventional search engines and therefore don’t appear in search results.

Why is it not indexed by typical search engines?
Some content can't be found through URL traversal

- Dynamic web pages and searchable databases
  - Response to a query or accessed only through a form
- Unlinked contents
  - Pages without any backlinks
- Private web
  - Sites requiring registration and login
- Limited access web
  - Sites with captchas, no-cache pragma http headers
- Scripted pages
  - Page produced by javascripts, Flash, etc.
In other times, content won’t be found

- **Crawling restrictions by site owner**
  - Use a robots.txt file to keep files off limits from spiders

- **Crawling restrictions by the search engine**
  - E.g.: a page may be found this way:
    http://www.website.com/cgi-bin/getpage.cgi?name=sitemap
  - Most search engines will not read past the ? in that URL

- **Limitations of the crawling engine**
  - E.g., real-time data – changes rapidly – too “fresh”
How big is Deep Web?

- Studies suggest it’s approx. 500x the surface Web
  - But cannot be determined accurately

- A 2001 study showed that 60 deep sites exceeded the size of the surface web (at that time) by 40x

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>URL</th>
<th>Web Size (GBs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>National Climatic Data Center (NOAA)</td>
<td>Public</td>
<td><a href="http://www.ncdc.noaa.gov/ol/satellite/satelliteresources.html">http://www.ncdc.noaa.gov/ol/satellite/satelliteresources.html</a></td>
<td>366,000</td>
</tr>
<tr>
<td>Right-to-Know Network (RTK Net)</td>
<td>Public</td>
<td><a href="http://www.rtk.net/">http://www.rtk.net/</a></td>
<td>14,640</td>
</tr>
<tr>
<td>MP3.com</td>
<td>Public</td>
<td><a href="http://www.mp3.com/">http://www.mp3.com/</a></td>
<td>4,300</td>
</tr>
<tr>
<td>Terraserver</td>
<td>Public/Fee</td>
<td><a href="http://terraserver.microsoft.com/">http://terraserver.microsoft.com/</a></td>
<td>4,270</td>
</tr>
</tbody>
</table>
Back in 2001, biggest fraction goes to databases.
Approaches for finding content in Deep Web

1. Specialized search engines

2. Directories
Specialized search engines

- **Crawl deeper**
  - Go beyond top page, or homepage

- **Crawl focused**
  - Choose sources to spider—topical sites only

- **Crawl informed**
  - Indexing based on knowledge of the specific subject
Specialized search engines abound

- There’s hundreds of specialized search engines for almost every topic
Directories

- Collections of pre-screened web-sites into categories based on a controlled ontology
  - Including access to content in databases

- **Ontology**: classification of human knowledge into topics, similar to traditional library catalogs

- Two maintenance models: **open** or **closed**
  - Closed model: paid editors; quality control (Yahoo)
  - Open model: volunteer editors; (Open Directory Project)
Ontologies allow for adding structure to Web content.
A particularly interesting search engine

Shodan lets the user find specific types of computers connected to the internet using a variety of filters

- Routers, servers, traffic lights, security cameras, home heating systems
- Control systems for water parks, gas stations, water plants, power grids, nuclear power plants and particle-accelerating cyclotrons

Why is it interesting?

- Many devices use "admin" as user name and "1234" as password, and the only software required to connect them is a web browser
How does Shodan work?

“Google crawls URLs – I don’t do that at all. The only thing I do is **randomly pick an IP** out of all the IPs that exist, whether it’s online or not being used, and I **try to connect** to it on different ports. It’s probably not a part of the visible web in the sense that you can’t just use a browser. It’s not something that most people can easily discover, just because it’s not visual in the same way a website is.”

John Matherly, Shodan's creator

- Shodan collects data mostly on HTTP servers (port 80)
  - But also from FTP (21), SSH (22) Telnet (23), and SNMP (161)
One can see through the eye of a webcam
Play with the controls for a water treatment facility
Find the creepiest stuff…

- Controls for a crematorium; accessible from your computer
No words needed

- Controls of Caterpillar trucks connected to the Internet

VIMS Onboard Time 2013/09/05 11:13:55
A Deep Web’s particular case

Dark Web
Dark Web

- **Dark Web** is the Web content that exists on darknets

- **Darknets** are overlay nets which use the public Internet but require specific SW or authorization to access
  - Delivered over small peer-to-peer networks
  - As hidden services on top of Tor

- The Dark Web forms a small part of the Deep Web, the part of the Web not indexed by search engines
The Dark Web is a haven for criminal activities

- Hacking services
- Fraud and fraud services
- Markets for illegal products
- Hitmen

...
Surface Web vs. Deep Web

**Surface Web**
- Size: Estimated to be 8+ billion (Google) to 45 billion (About.com) web pages
- Static, crawlable web pages
- Large amounts of unfiltered information
- Limited to what is easily found by search engines

**Deep Web**
- Size: Estimated to be 5 to 500x larger (BrightPlanet)
- Dynamically generated content that lives inside databases
- High-quality, managed, subject-specific content
- Growing faster than surface web (BrightPlanet)
Conclusions

- The Web is a major source of information for both criminal and legal investigation activities.

- The Web content that is typically accessible through conventional search engines is named the Surface Web and represents only a small fraction of the whole Web.

- The Deep Web includes the largest bulk of the Web, a small part of it (the Dark Web), being used specifically for carrying out criminal activities.
Primary bibliography

- Michael K. Bergman, *The Deep Web: Surfacing Hidden Value*
  
Flow analysis and intrusion detection