Studying the web - Scraping the open and dark web, crypto-markets, and discussion forums for intelligence

Dr. Richard Frank
rfrank@sfu.ca
Research Themes

CE Websites

Violent Extremist Websites

Violent Extremist Forums

Hacker Forums

Hacker Communities

Dark Web

CE Longitudinal

CE Geolocation


Cyber Security Cooperation Program
Canadian Internet Regulatory Agency
Canadian Safety and Security Program
Defence Research and Development Canada

CE = Child Exploitation
• Challenges
• Data Scraping Overview

• Data Sources
  • Websites
  • Online Discussion Forums
  • Social Media
  • Dark Web

• Conclusions

Unique scraping challenges
Processing
Analysis Example
Challenges

• There is an unmanageable amount of *data* on the Internet, all of which could contain actionable *information*.

• Manual searches and analysis can be conducted → insufficient.

• The possibilities increase vastly when advanced analytics can be applied to large datasets.
• We can easily do keyword searches

Find all posts with keywords ‘gun’ ‘shoot’ ‘school’

This leads to many false positives, ex: “Stop gun shootings at schools!”

• With advanced tools, we can instead ask

“Who should be a target of an intervention, based on their comments becoming more radical over time on the topic of shootings”

• By having entire discussion forums available, we can apply advanced analytics to detect patterns, networks of users, opinion leaders, or people on the path of becoming radicalized
DATA COLLECTION
• Goal: collect data and transform to intelligence

• There are multiple types of websites
  • Web 1.0
  • Discussion Forums
  • Social Media (Web 2.0)
  • Dark Web

• Each has different possibilities
• Each requires slightly different strategies
Challenges – Focus

• Challenge
  • Be a good citizen in the web
    • The Crawler should not impose too much of a burden on the web sites they crawl
    • Without proper ordering of the crawler activities, the crawler might unintentionally carry out a denial-of-service attack
  • The site might also block us, making data collection impossible

• Mitigations
  • Only create two requests/connections to each website
  • Use multiple computers, each with a different VPN exit node
Challenges – Scalability

• Challenge
  • The web is very large and evolving every day
  • Need a broad coverage and freshness → extremely high throughput

• Mitigation
  • With each instance limited to 2 simultaneous retrievals, need distributed crawling
    • Each crawler node is responsible for a subset of the URLs on the target website
Possible Solutions

• Off-the-shelf software is often not suitable
  • Hammers the target server with many connections
  • Cannot be distributed across many machines
  • Cannot scale to large websites
Our Infrastructure
The Dark Crawler

- Identifies Child Exploitation Material through MD5, SHA1, PhotoDNA hashes
- Current infrastructure: 1 web-server, 1 database server, 4 crawler machines
- Uses Virtual Private Networks to protect itself
- Only 2 connections per server allowed, to avoid being blocked
- Searches the clear-web and dark-web
- Geographically locates servers
- Uses WhoIS registry to collect information about the website owners

https://www.thedarkcrawler.com
EXAMPLE

• LECENing places to hide: Geo-Mapping Child Exploitation Material
Goals & Challenges

• Which sites are CE images located on? How can we find them?

• Automatically identifying images

• Where are those web-sites?

• Who do they belong to?
Goals & Challenges

- Which sites are CE images located on? How can we find them?

- Automatically identifying images

- Where are those web-sites?

- Who do they belong to?

- Searching on the ‘net randomly is useless
- 99.9999% of content is not CEM
- Guide the data collection through criteria, for example:
  - Keywords
  - Known images
  - Safe domains

<table>
<thead>
<tr>
<th>little</th>
<th>pthc</th>
</tr>
</thead>
<tbody>
<tr>
<td>boy</td>
<td>babyshivid</td>
</tr>
<tr>
<td>reelkiddymov</td>
<td>lover</td>
</tr>
<tr>
<td>mafiase</td>
<td>kiddy</td>
</tr>
<tr>
<td>Bibcamer</td>
<td>lolli</td>
</tr>
<tr>
<td>naked</td>
<td>qqaazz</td>
</tr>
</tbody>
</table>
Goals & Challenges

• Which sites are CE images located on? How can we find them?

• *Automatically identifying images*

• Where are those web-sites?

• Who do they belong to?

• Hash acts as a *fingerprint* to the image

• Each image is retrieved, hashed, hash value stored and compared to DB
  • No original images are stored
  • Location of CEM images is identified

<table>
<thead>
<tr>
<th>Cat</th>
<th>Count</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>702,997</td>
<td>Child exploitation. CCC s. 163.1(1)</td>
</tr>
<tr>
<td>2</td>
<td>2,109,813</td>
<td>“Gray-area”: An investigator was unable to determine if image depicts a person under the age of 18</td>
</tr>
<tr>
<td>3</td>
<td>49,419,190</td>
<td>“Other” material, not child exploitation</td>
</tr>
</tbody>
</table>
• Which sites are CE images located on? How can we find them?

• Automatically identifying images

• Where are those web-sites?

• Who do they belong to?

• Domains can easily be translated to IP addresses, ex:
  www.example.com → 123.45.56.78

• IP Addresses can be (with good certainty) translated to Long/Lat or City/State/Country
  123.45.56.78 → New York, NY, US

• Process known as geo-location
• Databases exist which provide this info
Goals & Challenges

• Which sites are CE images located on? How can we find them?

• Automatically identifying images

• Where are those web sites?

• Who do they belong to?

  • Each domain must be registered
  
  • Registrant information available in online databases (for ex: WhoISXML)

        www.example.com →

        John Doe, 2 Main St, NY, NY, US

  • This information changes as domains are bought/sold

  • Information can be anonymous

  • Is not applicable to blogs, etc, who use the domain name of the service provider
Our Infrastructure
The Dark Crawler

- Uses Virtual Private Networks to protect itself
- Only 2 connections per server allowed, to avoid being blocked
- Current infrastructure:
  - 1 web-server
  - 1 database server
  - 4 crawler machines

- Identifies Child Exploitation Material through MD5, SHA1, PhotoDNA hashes
- Searches the clear-web and dark-web

- Geographically locates servers
- Uses WhoIS registry to collect information about the website owners

https://www.thedarkcrawler.com
• Let crawler start from seed sites, and go until no CE content found
• Retrieved 54,156 web-pages & all images on them

• Found 4,937 Category 1&2 images

<table>
<thead>
<tr>
<th>Category</th>
<th>Images</th>
<th>Domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: CE</td>
<td>1,569</td>
<td>25</td>
</tr>
<tr>
<td>2: ‘gray area’</td>
<td>3,368</td>
<td>47</td>
</tr>
<tr>
<td>3: Unrelated</td>
<td>104,378</td>
<td>479</td>
</tr>
<tr>
<td>Not in Database</td>
<td>1,259,496</td>
<td>2,402</td>
</tr>
</tbody>
</table>

• Images spread across
  • 1,166 webpages
  • 63 servers/domains

• Encountered 29,272 videos (which we don’t deal with)
• Metadata generated by camera/software

• EXIF data was generally blank

1 image was taken with FinePix A345 [others had no such metadata]

10 were saved from Adobe Photoshop CS

1337 were saved from Picasa 3.0

No X/Y coordinates were stored
<table>
<thead>
<tr>
<th>Domain</th>
<th>Cat 1</th>
<th>Cat 2</th>
<th>Cat 1&amp;2</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www3.bfcollection.com">http://www3.bfcollection.com</a></td>
<td>0</td>
<td>1,590</td>
<td>1,590</td>
</tr>
<tr>
<td><a href="http://www.gayteenboylinks.com">http://www.gayteenboylinks.com</a></td>
<td>583</td>
<td>492</td>
<td>1,075</td>
</tr>
<tr>
<td><a href="http://www.pichack.com">http://www.pichack.com</a></td>
<td>438</td>
<td>369</td>
<td>807</td>
</tr>
<tr>
<td><a href="http://www.gayyoungboys.net">http://www.gayyoungboys.net</a></td>
<td>2</td>
<td>337</td>
<td>339</td>
</tr>
<tr>
<td><a href="http://www.gayteenlove.com">http://www.gayteenlove.com</a></td>
<td>180</td>
<td>0</td>
<td>180</td>
</tr>
<tr>
<td><a href="http://www.gayteenmovies.net">http://www.gayteenmovies.net</a></td>
<td>6</td>
<td>148</td>
<td>154</td>
</tr>
<tr>
<td><a href="http://www.adult-empire.com">http://www.adult-empire.com</a></td>
<td>56</td>
<td>25</td>
<td>81</td>
</tr>
<tr>
<td><a href="http://www.twinksecrets.com">http://www.twinksecrets.com</a></td>
<td>2</td>
<td>51</td>
<td>53</td>
</tr>
<tr>
<td><a href="http://www.allcuteboys.com">http://www.allcuteboys.com</a></td>
<td>51</td>
<td>0</td>
<td>51</td>
</tr>
<tr>
<td><a href="http://wetmike.com">http://wetmike.com</a></td>
<td>20</td>
<td>28</td>
<td>48</td>
</tr>
<tr>
<td><a href="http://cuteboyblog.com">http://cuteboyblog.com</a></td>
<td>36</td>
<td>9</td>
<td>45</td>
</tr>
<tr>
<td><a href="http://thegayteentube.com">http://thegayteentube.com</a></td>
<td>19</td>
<td>16</td>
<td>35</td>
</tr>
<tr>
<td><a href="http://teengaysvideo.com">http://teengaysvideo.com</a></td>
<td>19</td>
<td>16</td>
<td>35</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Websites with Category 1 & 2

Translate to server address(es)

GeoDB

X/Y coordinate
<table>
<thead>
<tr>
<th>Domain</th>
<th># IPs</th>
<th># of Pages</th>
<th>Cat1</th>
<th>Cat2</th>
<th>Cat3</th>
<th>CatUnk</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://xboysvideos.com:80">http://xboysvideos.com:80</a></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>22</td>
<td>68</td>
</tr>
<tr>
<td><a href="http://boywinktube.com:80">http://boywinktube.com:80</a></td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>20</td>
<td>44</td>
</tr>
<tr>
<td><a href="http://susodroika.info:80">http://susodroika.info:80</a></td>
<td>1</td>
<td>5</td>
<td>16</td>
<td>10</td>
<td>55</td>
<td>125</td>
</tr>
<tr>
<td><a href="http://www2.boysvideos.com:80">http://www2.boysvideos.com:80</a></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>21</td>
<td>70</td>
</tr>
<tr>
<td><a href="http://www.twinksexcrets.com:80">http://www.twinksexcrets.com:80</a></td>
<td>1</td>
<td>8</td>
<td>2</td>
<td>51</td>
<td>175</td>
<td>1283</td>
</tr>
</tbody>
</table>

...more...
- Content accessed from Canada
- Website hosted in Netherlands
- Domain belongs to Michail in Russia
How to attack this?

- Arrest Michail?
  - He’s in Russia! 😊

- Shut server down?
  - 2 problems:
    1) It’s easy to replace, rent another and re-upload content
    2) Content is *not* on suxodro4ka.info
suxodro4ka.info *references* images from

<table>
<thead>
<tr>
<th>Domain</th>
<th>CE</th>
<th>Gray</th>
<th>Unrelated</th>
<th>??</th>
<th>Country</th>
<th>Registrant</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://img.clicksagent.com">http://img.clicksagent.com</a></td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td>Russia</td>
<td>Michail</td>
</tr>
<tr>
<td><a href="http://promo.adultempire.com">http://promo.adultempire.com</a></td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td>Cyprus</td>
<td></td>
</tr>
<tr>
<td><a href="http://suxodro4ka.info">http://suxodro4ka.info</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Russia</td>
<td>Michail</td>
</tr>
<tr>
<td><a href="http://www.allcuteboys.com">http://www.allcuteboys.com</a></td>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td>England</td>
<td></td>
</tr>
<tr>
<td><a href="http://www.gayteenlove.com">http://www.gayteenlove.com</a></td>
<td>10</td>
<td>29</td>
<td></td>
<td></td>
<td>England</td>
<td>Anonymous</td>
</tr>
<tr>
<td><a href="http://www.pichack.com">http://www.pichack.com</a></td>
<td>12</td>
<td>10</td>
<td>44</td>
<td>34</td>
<td>England</td>
<td>England</td>
</tr>
<tr>
<td><strong>Grand Total</strong></td>
<td>16</td>
<td>10</td>
<td>73</td>
<td>183</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Content accessed from Canada
- Website hosted in Netherlands
- Registrant in Russia
- Website contains large number of images from site hosted in US, registered in UK
Where are the Servers?

• Image → Domain → Server Address → GeoDB → XY

• XY coordinate was reverse-geocoded to get address

• Servers containing CE material

<table>
<thead>
<tr>
<th>Country Name</th>
<th>Cat 1</th>
<th>Cat 2</th>
<th>Cat 1&amp;2</th>
</tr>
</thead>
<tbody>
<tr>
<td>??</td>
<td>1 (2.38%)</td>
<td>1 (2.27%)</td>
<td>1 (1.59%)</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>2 (4.76%)</td>
<td>3 (6.82%)</td>
<td>4 (6.35%)</td>
</tr>
<tr>
<td>United States</td>
<td>30 (71.43%)</td>
<td>31 (70.45%)</td>
<td>47 (74.60%)</td>
</tr>
<tr>
<td>Canada</td>
<td>1 (2.38%)</td>
<td>1 (2.27%)</td>
<td>1 (1.59%)</td>
</tr>
<tr>
<td>Dominica</td>
<td>1 (2.38%)</td>
<td>1 (2.27%)</td>
<td>1 (1.59%)</td>
</tr>
<tr>
<td>Netherlands</td>
<td>7 (16.67%)</td>
<td>7 (15.91%)</td>
<td>9 (14.29%)</td>
</tr>
</tbody>
</table>

Total: 42 44 63
### Distribution of Contact Info

<table>
<thead>
<tr>
<th>Country</th>
<th>Anon</th>
<th>Name Given</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Australia</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Austria</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Bahamas</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Belarus</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Cyprus</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Czech Republic</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Estonia</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Israel</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Italy</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Luxembourg</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Mexico</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Netherlands</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Panama</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>Russian Federation</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Sweden</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>United Kingdom</td>
<td>52</td>
<td>15</td>
</tr>
<tr>
<td>United States</td>
<td>62</td>
<td>12</td>
</tr>
<tr>
<td><strong>Grand Total</strong></td>
<td>144</td>
<td>94</td>
</tr>
</tbody>
</table>
DISCUSSION FORUMS
Challenges

- Regular Webpage
  - Unstructured
  - Hard to automatically understand
Challenges

- Forums
- Structured data
- Can exploit structure to better understand content
There are standard languages for getting information from trees (which HTML is): XPath

There are standard languages for cleaning up text: XQuery

1827
Information from Discussion Forums

- Using XPaths and Xqueries, we parse apart entire discussion forums

- **Sub-Forums**
  - Names
  - Descriptions

- **Threads**
  - Thread IDs
  - Title
  - Start Date
  - # of Views
  - # of Replies…

- **Posts**
  - Author
  - Join Date
  - Rank
  - Post Content
  - Post Date
  - …
~65m posts currently captured
Forum capture includes all sub-forums, all threads and all posts.
Forum capture includes all sub-forums, all threads and all posts.
We are building the analysis tools for looking at users.
Join Date of Users

Join Date (Month and Year)
Subject: Refutation of Rabee al-Madkhali: Architect of the "Jarh wal-Ta'dil" Cult
Number of Posts: 1768
OP: justabro

Size of node = total posts across entire forum
Size of tie = number of replies one author made after the another
Identification of radical authors in online environments
• Develop novel computational techniques and algorithms

• Develop a technique to identify radical authors in online discussion forums

• Find the more radical and infectious threads, authors, postings, ideas and/or ideologies
- Take a discussion forum (or 4)
- Extract all the high-frequency keywords
- Can use open-source software such as OpenNLP, or CoreNLP to do this
Within each message, determine sentiment around keyword
Assigns polarity values to those keywords

<table>
<thead>
<tr>
<th>Message ID</th>
<th>Keyword 1</th>
<th>Keyword 2</th>
<th>...</th>
<th>Keyword 67</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Author A - Post 1</th>
<th>-2</th>
<th>1</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post 2</td>
<td>3</td>
<td>-4</td>
<td></td>
</tr>
<tr>
<td>Post 3</td>
<td>0.5</td>
<td>-5</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post 99</td>
<td></td>
<td></td>
<td>-31</td>
</tr>
</tbody>
</table>

STEP I
Forum Data

STEP II
Parts-of-Speech Tagging

STEP III
Sentiment Analysis

STEP IV
SIRA

STEP V
Radical Score
**SIRA Algorithm:**

1.) Relative average sentiment score (10 points)

2.) Volume of negative posts (10 points)

3.) Severity of negative posts (10 points)

4.) Duration of negative posts (10 points)

---

**RADICAL SCORE**

(40 points)

<table>
<thead>
<tr>
<th>Author A – Post</th>
<th>Keyword 1</th>
<th>Keyword 2</th>
<th>...</th>
<th>Keyword 67</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post 1</td>
<td>-2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post 2</td>
<td>3</td>
<td>-4</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>Post 3</td>
<td>0.5</td>
<td>-5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post 99</td>
<td></td>
<td></td>
<td></td>
<td>-31</td>
</tr>
</tbody>
</table>
• Four Islamic-based web-forums
• ~1 million posts made by ~26,000 unique authors

Results

Radical Score

Number of Authors (Frequency)

Distribution of Radical Scores Across Sample

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>18.049</td>
</tr>
<tr>
<td>SD</td>
<td>8.018</td>
</tr>
<tr>
<td>N</td>
<td>25,869</td>
</tr>
</tbody>
</table>
## Results

Table 1. Most Radical Users Across Four Islamic Discussion Forums

<table>
<thead>
<tr>
<th>Author</th>
<th>Web-Forum</th>
<th>Volume of Posts</th>
<th>Average Posting Score</th>
<th>Negative Posts (%)</th>
<th>Very Negative Posts (%)</th>
<th>Duration of Negative Posts (Years)</th>
<th>Radical Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prism</td>
<td>Islamic Awakening</td>
<td>139</td>
<td>-4.47</td>
<td>97.1</td>
<td>23.9</td>
<td>2.87</td>
<td>39.028</td>
</tr>
<tr>
<td>O.M.T.</td>
<td>Gawaher</td>
<td>2,874</td>
<td>-1.02</td>
<td>72.4</td>
<td>0.6</td>
<td>6.69</td>
<td>38.934</td>
</tr>
<tr>
<td>Nash</td>
<td>Islamic Awakening</td>
<td>425</td>
<td>-1.05</td>
<td>70.3</td>
<td>1.2</td>
<td>5.69</td>
<td>38.876</td>
</tr>
<tr>
<td>Ray</td>
<td>Gawaher</td>
<td>524</td>
<td>-1.29</td>
<td>75</td>
<td>1.1</td>
<td>3.35</td>
<td>38.823</td>
</tr>
<tr>
<td>3D</td>
<td>Gawaher</td>
<td>28</td>
<td>-2.21</td>
<td>85.7</td>
<td>3.6</td>
<td>3.94</td>
<td>38.790</td>
</tr>
<tr>
<td>Harley</td>
<td>Gawaher</td>
<td>1525</td>
<td>-0.88</td>
<td>64.0</td>
<td>0.3</td>
<td>6.28</td>
<td>38.652</td>
</tr>
<tr>
<td>JHU</td>
<td>Islamic Awakening</td>
<td>423</td>
<td>-1.27</td>
<td>74.1</td>
<td>0.7</td>
<td>2.76</td>
<td>38.625</td>
</tr>
<tr>
<td>Lee</td>
<td>Islamic Awakening</td>
<td>2770</td>
<td>-1.47</td>
<td>76.8</td>
<td>0.7</td>
<td>2.26</td>
<td>38.620</td>
</tr>
</tbody>
</table>
## Results

Muslims are the current object of our society's fear and contempt....

The aggressive American war machine continues to endeavour tirelessly to [...] convey a message of truth to others without distortion or perversion.
...it dont matter which non muslim country gonna respond. the chain reaction gonna lead to doomsday [sic]

...the us [United States] loves to display its donkey talent in poking its nosey nose into others' affairs [...] merely to satisfy the us' [United States] lust for revenge...[sic]
Potential trajectory groups for discussions on Jews in Stormfront
EXAMPLE

• Influential hackers from 12 hacking forums
<table>
<thead>
<tr>
<th>Forum ID</th>
<th>Number of Posts</th>
<th># of Authors</th>
<th># of Threads</th>
<th>First Post</th>
<th>Last Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>935,449</td>
<td>65,933</td>
<td>36,468</td>
<td>4/7/2011</td>
<td>4/23/2018</td>
</tr>
<tr>
<td>11</td>
<td>1,145,378</td>
<td>49,856</td>
<td>24,574</td>
<td>4/8/2013</td>
<td>4/12/2018</td>
</tr>
<tr>
<td>158</td>
<td>2,442,669</td>
<td>10,230</td>
<td>170,643</td>
<td>5/30/2016</td>
<td>11/5/2017</td>
</tr>
<tr>
<td>159</td>
<td>1,581,238</td>
<td>33,534</td>
<td>112,436</td>
<td>2/9/2002</td>
<td>4/8/2018</td>
</tr>
<tr>
<td>160</td>
<td>263,892</td>
<td>9,497</td>
<td>32,999</td>
<td></td>
<td></td>
</tr>
<tr>
<td>161</td>
<td>110,037</td>
<td>9,132</td>
<td>13,265</td>
<td>1/2/2016</td>
<td>4/19/2018</td>
</tr>
<tr>
<td>162</td>
<td>185,690</td>
<td>5,615</td>
<td>24,343</td>
<td>9/13/2002</td>
<td>4/18/2018</td>
</tr>
<tr>
<td>165</td>
<td>246,035</td>
<td>16,106</td>
<td>28,180</td>
<td>8/1/2005</td>
<td>4/28/2018</td>
</tr>
<tr>
<td>175</td>
<td>542,399</td>
<td>13,168</td>
<td>58,151</td>
<td>6/25/2008</td>
<td>5/2/2018</td>
</tr>
</tbody>
</table>

- 12 forums captured in their publicly available entirety
  - **14.8 million** posts
  - **438,542** users
Analysis of Usernames

• We assumed that a username is consistent across forums

  same username =?= same user

• Realistic?

• We argue that in an anonymous environment, the reputation (trust) in a pseudonym is everything
  • There’s a cost to establishing trust in a new username
  • There’s a cost to maintaining many disparate usernames

• Which is better, lots of trust in a single username, or small amounts of trust in multiple usernames?
### Table: Users and the number of other forums they contribute to

<table>
<thead>
<tr>
<th>Forum ID</th>
<th>Number of users in n other forums</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n = 0$</td>
</tr>
<tr>
<td>10</td>
<td>54,210 (82.23%)</td>
</tr>
<tr>
<td>11</td>
<td>39,360 (78.95%)</td>
</tr>
<tr>
<td>157</td>
<td>204,092 (96.12%)</td>
</tr>
<tr>
<td>158</td>
<td>9,397 (91.86%)</td>
</tr>
<tr>
<td>159</td>
<td>29,445 (87.81%)</td>
</tr>
<tr>
<td>160</td>
<td>8,334 (87.77%)</td>
</tr>
<tr>
<td>161</td>
<td>7,000 (76.65%)</td>
</tr>
<tr>
<td>162</td>
<td>4,544 (80.94%)</td>
</tr>
<tr>
<td>164</td>
<td>11,412 (88.90%)</td>
</tr>
<tr>
<td>165</td>
<td>13,637 (84.68%)</td>
</tr>
<tr>
<td>167</td>
<td>175 (58.52%)</td>
</tr>
<tr>
<td>175</td>
<td>10,472 (79.59%)</td>
</tr>
</tbody>
</table>

**Total (overlap)**: 392,078 (89.41%) | 34,542 (7.87%) | 7,773 (1.77%) | 2,524 (0.57%) | 1,020 (0.23%) | 354 (0.08%) | 126 (0.02%) | 72 (0.01%) | 438,489

**Total (no overlap)**: 392,078 (94.96%) | 17,271 (4.18%) | 2,591 (0.62%) | 631 (0.15%) | 204 (0.04%) | 59 (0.01%) | 18 (0.00%) | 9 (0.00%) | 412,861

**Total meeting selection criteria**: 193 | 57 | 18 | 9 | 277

412,861 users across the 12 forums, after de-duplication
• Too many users → must sample

• For our sampling, an author must:
  • Be active in $\geq 5$ forums
  • Have no less than 10% of posts in one forum
  • Have no more than 85% of posts in one forum
  • Have 75% of the posts be in at least two forums

• Authors that satisfied criteria were selected for further qualitative analysis

• Resulting sample after applying selection criteria:
  • 195 users
  • 36,978 total posts
**Analysis: Author Diversification**

- Possible to identify changes in consumption of information by looking at a user’s activity *across* forums
  - *Fred*
    - Forum 10/11 – Consumer
    - Forum 158/159 – Contributor
  - *Ulfrid*
    - Forum 10/11/157 – Consumer
    - Forum 175 – Contributor

- *Fred* and *Ulfrid* may be moving content between forums
  - Providing more users with access to potentially malicious information

<table>
<thead>
<tr>
<th>Author</th>
<th>Forum</th>
<th>RQ%</th>
<th>THX%</th>
<th>SHR%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fred</td>
<td>10</td>
<td>25%</td>
<td>75%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>11%</td>
<td>81%</td>
<td>7%</td>
</tr>
<tr>
<td></td>
<td>157</td>
<td>50%</td>
<td>0%</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>158</td>
<td>20%</td>
<td>4%</td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>159</td>
<td>20%</td>
<td>10%</td>
<td>70%</td>
</tr>
<tr>
<td></td>
<td>164</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>Ulfrid</td>
<td>10</td>
<td>33%</td>
<td>67%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>32%</td>
<td>60%</td>
<td>8%</td>
</tr>
<tr>
<td></td>
<td>157</td>
<td>29%</td>
<td>53%</td>
<td>17%</td>
</tr>
<tr>
<td></td>
<td>159</td>
<td>82%</td>
<td>9%</td>
<td>9%</td>
</tr>
<tr>
<td></td>
<td>175</td>
<td>28%</td>
<td>9%</td>
<td>63%</td>
</tr>
</tbody>
</table>
Conclusions about the broader community

• The vast majority of these discussions are about general issues, only a small portion (9.3%) tend to share content

• ~95% of users on these forums tend to specialize into a single forum

• Only a few users seem to consume content on some, share content on other forums

• If they have reputation, then it is spread across multiple pseudonyms, significantly decreasing the value of each pseudonym
  • A lesser known user on one forum may be a key contributor/mentor on several others

• Hackers seem to put all eggs in a single basket → LE strategy of taking down forums works well in this environment → hackers lose all reputation, must rebuild after a takedown
These pages can be treated the same forums, due to the repeated structure & similar content.
DARK WEB
Welcome to the Internet

Size indicates popularity
Welcome to the Internet

Size indicates popularity
Welcome to the Internet

Some 46b webpages!!
Welcome to the Internet

Some 46b webpages!!

Some 1trillion webpages!!
(by many estimates)
From tumblr
Either use a proxy, or say hi to the police.

The Internet (more or less)

You can turn a computer on. Yay.
• A hidden part of the internet, helps to protect the privacy of internet users from traffic analysis attacks
• Recognized as an ideal hosting ground for many illegal activities and services, from large drug markets to child sexual abuse.

• History
  • initially developed by the U.S Navy in the 1990s
  • was launched officially in 2002 for the public with the core objective of open source information gathering

• Characteristics
  • Decentralized
  • Internet infrastructure utilization
  • Usage of non-standard computer protocols and ports unreachable by those outside the network
• Network of 50,000 computers
• Pick 3, create a path
• Each path is random
• Each path is encrypted multiple times
• Alice does not know where Bob is, nor Bob know where Alice is
• Paths are rebuilt every 10 minutes
Fentanyl

Selling 16.8mg = 119Euro
Credit Cards
Hacking
This wiki page discusses resources specifically for people who are attracted to children. This can include everything from discussion groups to ostensibly legal images of children in depictions of so-called child pornography. The term children here refers to children and teenagers.

Child pornography is illegal in most of the world, and most of the world seems to have different views on it. The aim of this page is to provide information on how to create and share this material. It is the source of constant controversy, but these pages do not contain any legal advice. The following projects are a group effort and will be used by the community to keep track of as many resources as possible.

Projects

- Age of consent - Information on the statutes and legal codes or conduct regulations.
- History of CP - An attempt to make an encyclopedia of Child Pornography.
- Child Models, MySpace girls and the like.
- Centralization Project - For new it is in preparation about creating a distributed wiki-like database complete with image samples and file hashes.
- To do - List of suggestions to reconcile society with child lovers.
- Your Own Pede Site - A tutorial on making a pedo site of your own, and replacing the anonymous security disks.
- Pedo tag - Discussion on how to identify the pedo-community.
- Back to Normal - Blueprint for a new child love movement. Volunteers sought! (Last Updated 2001-09-26)
- The Upshot Guide - Project to set a guide for ciamat uploads. Includes fast list.
- Child Nudity Wiki - An indexing of all instances of child nudity.

Chat / Discussion Groups

- BBS
  - GitChat - Board for chatting about girl love. Strongly anti-rape. Registration required to post. The hint of anything illegal will get you banned.
  - La Jardin d'Alice and La Maison de Sabrina - French girl-lover boards. The first is modeled after GitChat, the second is a phpb. Similar rules to GitChat: registration required, hinting at anything illegal warrants you an immediate ban. Discussing someone illegal you have done (including having watched child pornography) you have not been prosecuted for is prohibited.
  - Newgen - A site for the discussion of physical attractions and relationships between minors and adults.

Resources

- 4.1 Non-Hide
  - 4.1.1 WWW
    - 4.1.1.1 Linked Portals
    - 4.1.1.2 Communities
    - 4.1.1.3 Image Boards
    - 4.1.1.4 Image Indexes
    - 4.1.1.5 Torrent Sites
- 4.2 Nude or Hardcore
  - 4.2.1 Freestore
    - 4.2.1.1 Freestore
    - 4.2.1.2 Fast Boards
    - 4.2.1.3 Software Download
  - 4.2.2 IDP
    - 4.2.2.1
  - 4.2.3 Tor
    - 4.2.3.1 Web Hidden Services
    - 4.2.3.1.1 General Sites
    - 4.2.3.1.2 Boards
    - 4.2.3.1.3 Sites with Access Restrictions
    - 4.2.3.1.4 LS
    - 4.2.3.1.5 How-to Guides
    - 4.2.3.1.6 Test-Only
    - 4.2.3.2 Dual Web Hidden Services
    - 4.2.3.3 TorChat EDS
- 4.2.4 WWW
  - 4.2.4.1 Sample Search Engine Queries
  - 4.2.5 Combination
  - 4.2.5.1 LS

5 Mirrors
Guns
More Drugs

SW Asian #4 Heroin

The Heroin we offer comes **direct from the importer** with no middle man.

It is white/light beige in color and we take great pride in the fact that **we do not cut our product whatsoever** and we ensure that our source does not do so either.

We would much prefer to offer a high quality product and have repeat customers.

What's the difference between #3 and #4 heroin?

Most people know this but we get this asked often enough that we decided to add this information onto the listing.

The most commonly found forms of Heroin are #3 and #4.

No.4 grade Heroin is the **PUREST** form of Heroin. It is a white powder that is easily dissolved in water and thus readily injected.

No.3 grade Heroin is a tan-coloured, granular product often referred to as “brown rock.”

People’s Drug Store pride ourselves on the **HIGH QUALITY** of our amazing, **HIGH PURITY COCAINE**.

Just like our Heroin, we get our cocaine **DIRECT** from the importer (usually from Peru but sometime its from Columbia) and we always get our product given to us right off of the kilogram bricks as they come in so we can be **ABSOLUTELY POSITIVE** that it was not cut or

In locally in any way!!

Policy to **NEVER** cut or add any kind of adulterant to any of our products.

Rest assured that any cocaine you buy from us will be as pure as it was on the boat from South America to cross the ocean.

<table>
<thead>
<tr>
<th>Price</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>22 USD = 0.105</td>
<td>1 X</td>
</tr>
<tr>
<td>37 USD = 0.176</td>
<td>1 X</td>
</tr>
</tbody>
</table>

Wanna Make Some Free BTC??

Tell others about this shop, and earn 5% from every purchase they will make. Simply give them the following link:
http://www.peoplesdrugstore.org/?ref=YOURUSERNAME (or the original http://newdrugstoremarket.onion/?ref=YOURUSERNAME)
Replace YOURUSERNAME with your actual username on the site and get earnings directly to your wallet.

Cocaine(85%) & Crack
Welcome to AppleTor

We have the latest and greatest Apple products at wholesale prices!

Learn more

Cheap
We’re cheap. That’s that. We can buy Apple products at wholesale prices.

And... Shipping is free!

Reliable
Unlike other services on Tor. You can rely on us!

We respond to every email within 24 hours.

Fast
We guarantee you’ll get your purchase in 7 business days. Around the globe.
FRESH CC/CVV FROM USA VISA/MASTER/DISCOVER Cards

Price: $14.00 USD / 0.062235 BTC

Place Order

Price: $14.00 USD / 0.062235 BTC

Vendor: 609austin14 +USB advanced

Transaction: Auto-FE

Ships From: Unknown Anime Land

Ships To: Worldwide

Delivery Time (Days): 1

Category: Money
5 gr Dry Amphetamine-Speed *Dry

Price: $27 USD / 0.120026 BTC

Vendor: love2xlr8 =488 =advanced
Transaction: Auto-FE
Ships From: Holland
Ships To: Worldwide
Delivery Time (Days): 10
Category: Speed
<table>
<thead>
<tr>
<th>Model</th>
<th>Price</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desert Eagle 357 Mag</td>
<td>~$2,435CAD</td>
<td>.357 Magnum, Titanium Gold, Barrel Length 5.5 in, Capacity 8, Semi-Auto</td>
</tr>
<tr>
<td>Remington Defense XM110 SASS 308</td>
<td>~$19,388CAD</td>
<td>7.62x51mm NATO, Gas operated, 20 rounds, Length 1028 mm, Weight 3.440 kg</td>
</tr>
<tr>
<td>Barrett M107A1 20&quot; CQ FDE 50 BMG CDL Suppressor</td>
<td>~$2,435CAD</td>
<td>7.62x51mm NATO, Gas operated, 20 rounds, Length 1028 mm, Weight 3.440 kg</td>
</tr>
</tbody>
</table>
Hackers

Prices:
I'm not doing this to make a few bucks here and there. I'm not from some crappily ran europe country and happy to scam people for 50 euro.
I'm a professional computer expert who could earn 50-100 euro an hour with a legal job.
So stop reading if you don't have a serious problem worth spending some cash at.
Prices depend a lot on the problem you want me to solve, but minimum amount for smaller jobs is 200 euro.
You can pay me anonymously using Bitcoin.

Technical skills:
- Web (HTML, PHP, SQL, APACHE)
- C/C++, Assembler, Delphi
- 0day Exploits, Highly personalized trojans, Bots, DDOS
- Spear Phishing Attacks to get accounts from selected targets
- Basically anything a hacker needs to be successful, if I don't know it, I'll learn it very fast.
- Anonymity: no one will ever find out who I am.

Social Engineering skills:
- Very good written and spoken (phone calls) English and German.
- If I can't hack something technically, I'll make phone calls or write emails to the target to get the needed information, I have had people make things you wouldn't believe really often.
- A lot of experience with security practices inside big corporations.

What I'll do:
I'll do anything for money, I'm not a pussy :) if you want me to destroy some business or a person's life, I'll do it!
Some examples:
Simply hacking something technically
Causing a lot of technical trouble on websites/networks to disrupt their service with DDOS and other methods.
Economic espionage
Getting private information from someone
Ruining your opponents, business or private persons you don't like, I can ruin them financially and get them arrested, whatever you like.
If you want someone to get known as a child porn user, no problem.
Hosting

- TorWeb - The largest hosting company on TOR. Offering webhosting and VPS packages
- Secured Hosting - Hidden Service Hosting with SSH login.
- Web Hosting - Web Hosting - PHP5, MySQL, SFTP Access, .onion Domain. 24 hours free hosting.
- TorShops - Get your own .onion store with full bitcoin integration.
- Img Hosting - Host and sell your original pictures for Bitcoins.
- Liberty's Hackers - Service and Hosting Provider in onionland
- Liberty's Hackers - new address for the liberty's hackers hidden service, same conditions that above.
- CYRUSERV - Hosting service with an emphasis on security, open for business again.
- Onionweb filehosting - Filehosting service. 100MB upload limit, no illegal files allowed.
Brand-specific

03/11/2015: Christmas sale begins!

Terms and conditions:
Price: 50% discount, all actual Samsung models available
Check: amazon.co.uk, samsung.co.uk
Express shipping with tracking
Welcome to Fake Documents Service - the unique producer of quality fake documents. We offer only original high-quality fake passports, driver's licenses, ID cards, stamps and other products for following countries: Australia, Belgium, Brazil, Canada, Finland, France, Germany, Italy, Netherlands, UK, USA and some others.

If you want to learn more about what kinds of documents can be found in our website please visit the sections "Services" and "Samples". You can find more details about ordering procedure and additional details visiting the sections "FAQ" and "Order".
Each capture period persisted for 8 days and was seeded using URLs from a previous study.

<table>
<thead>
<tr>
<th>Date</th>
<th>Crawl ID (Wave number)</th>
<th>Pages Found but Not Retrieved</th>
<th>Pages Retrieved</th>
<th>Domains Sampled</th>
</tr>
</thead>
<tbody>
<tr>
<td>April 24, 2016</td>
<td>50(1)</td>
<td>3,832,608</td>
<td>1,833,916</td>
<td>6,282</td>
</tr>
<tr>
<td>October 7, 2017</td>
<td>157(2)</td>
<td>3,145,799</td>
<td>534,894</td>
<td>1,105</td>
</tr>
<tr>
<td>December 23, 2017</td>
<td>162(3)</td>
<td>209,393</td>
<td>22,162</td>
<td>4,459</td>
</tr>
<tr>
<td>January 1, 2017</td>
<td>171 (4)</td>
<td>238,495</td>
<td>21,957</td>
<td>2,393</td>
</tr>
<tr>
<td>January 9, 2017</td>
<td>173 (5)</td>
<td>125,390</td>
<td>13,099</td>
<td>2,303</td>
</tr>
<tr>
<td>January 17, 2017</td>
<td>174 (6)</td>
<td>296,132</td>
<td>32,774</td>
<td>5,271</td>
</tr>
<tr>
<td>January 27, 2017</td>
<td>175 (7)</td>
<td>1,033,144</td>
<td>120,983</td>
<td>5,496</td>
</tr>
<tr>
<td>February 6, 2017</td>
<td>176 (8)</td>
<td>74,661</td>
<td>5,654</td>
<td>2,255</td>
</tr>
<tr>
<td>February 15, 2017</td>
<td>177 (9)</td>
<td>948,395</td>
<td>97,697</td>
<td>3,912</td>
</tr>
<tr>
<td>February 27, 2017</td>
<td>179 (10)</td>
<td>397,701</td>
<td>30,451</td>
<td>3,533</td>
</tr>
<tr>
<td>March 7, 2017</td>
<td>180 (11)</td>
<td>983,121</td>
<td>92,780</td>
<td>4,857</td>
</tr>
<tr>
<td>March 21, 2017</td>
<td>181 (12)</td>
<td>291,749</td>
<td>24,269</td>
<td>4,269</td>
</tr>
</tbody>
</table>
Manually Categorized 774 Sites

- Total of 11,014 unique domains were found
- 774 were randomly sampled for manual classification and analysis
Manually Categorized 774 Sites

LEGALITY OF DARK WEB SITES

- Illegal: 52%
- Legal: 48%

Ambiguous: 2 sites (0%)
Manually Categorized 774 Sites

Focus of sites in the Dark Web

- CE
- Counterfeit
- Directory
- Down
- Drugs
- eCurrency
- FileSharing
- Gambling
- Hacking
- Hitman
- Hosting
- Mail
- Misc
- Pornography
- Privacy
- Privacy
- Privacy
- (blank)

Number of Sites

- CE: 20
- Counterfeit: 1
- Directory: 1
- Down: 1
- Drugs: 1
- eCurrency: 1
- FileSharing: 1
- Gambling: 1
- Hacking: 1
- Hitman: 1
- Hosting: 1
- Mail: 1
- Misc: 2
- Pornography: 1
- Privacy: 1
- Privacy: 1
- Privacy: 1
- (blank): 0
Selling illicit drugs online – why?
   Safer
   Global reach
   More convenient

**Silk Road** was the most infamous example of this trend
   Emphasized “responsible vendors” and “intelligent consumers”
   When taken down, $1.2 billion in sales revenue (FBI)
• Started with Yandex, searched for “24 biz”

• Found 47 sites, which were manually looked at

• 28 sites contained data, others empty or unavailable
• Websites shared one of 4 templates
• 4 rules for ad summary
• 4 rules for ad details
• Each piece contains
  • Website
  • City, district
  • Drug name
  • Drug price
  • Amount in grams
  • Substance type (e.g. cannabinoids or stimulants)

935 data pieces
Distribution of Drug Types Among Marketplaces

Substance Types

- ATS
- Hallucinogens
- Natural cannabinoids
- Synthetic cannabinoids

Marketplaces

- 2add.biz
- mr24.biz
- chem4.biz
- 28p.biz
- sc24.biz
- nicehop24.biz
- wshop24.net
- urshop24.biz
- nar24.biz
- speedliner24.biz
- 240qg.biz
- exshop24.biz
- karmav24.klar
- bit
- kmv24.biz
- bioticbye-biz
- bro24.biz
- ek24.biz
- uckop24.biz
- bounty24.biz
- popeye24.biz
- pandor24.biz
- tran24.biz
- mkp24.biz
- b2d21.biz
- zpc24.biz
- zdey24.biz

Drug Amount, Grams

- 300
- 250
- 200
- 150
- 100
- 50
- 0
*ATS = amphetamine-type stimulants
Payment Methods
• By collecting *targeted* data for *specific purposes*, you can change the questions you ask

• **Steps**
  • Create a *good* question 😊
  • Pick websites/forums/social media/etc where this information exists
  • Collect data
    • Off-the-shelf crawlers [slow 😞]
    • Custom-built crawlers [fast, but more effort]
  • Extract relevant data using *XPath* and *XQuery* [standard languages any programmer can implement]
  • Apply SNA, sentiment, keyword analysis, geolocation, etc, to extracted data
  • Look for patterns, pick your targets for in-depth analysis
Studying the web - Scraping the open and dark web, crypto-markets, and discussion forums for intelligence

Dr. Richard Frank
rfrank@sfu.ca