Preliminary Experiments on Measuring Web Censorship Around the World

Minaxi Gupta
School of Informatics and Computing
minaxi@cs.indiana.edu
http://www.cs.indiana.edu/~minaxi
Motivation

- Recent events in Egypt and neighboring countries
- Lack of ongoing projects to measure censorship in a technically sound way
Goals

- measure who censors what, how, when on an ongoing basis
- design practical anti-censorship evasion techniques
Censorship on the Web

- Prevention of access to specific Web content or hosts
  - Blocking websites (URLs, domains)
  - Blocking IPs
  - Blocking ports, protocols
  - Blocking keywords
  - Removing content (ex: Egypt)
  - Blocking access to the Web
Who censors?

- Much more prevalent than we would imagine
- Many countries perform censorship under various pretexts
- Examples include: Australia, Burma, Bahrain, China, Cuba, Egypt, Ethiopia, France, Iran, N. Korea, Russia, Saudi Arabia, S. Korea, Syria, Thailand, Tunisia, Turkmenistan, Uzbekistan, Vietnam
Current censorship measurement efforts

- Most are from journalists, using non-technical methods
- Technical projects:
  - Berkman center’s open net initiative (ONI): 2004, not currently active
  - Mao’s PAM 2011 work on measuring filtering points in China from the outside
  - Feamster’s USENIX 2010 work on anti-censorship system
  - kc’s work on measuring censorship-related outages
Our methodology

- Access websites, IPs, ports, protocols, keywords using free proxies within censoring countries
- Compare responses with accesses from the U.S. to infer censorship
- Simple idea but the devil lies in the details
Why drawing conclusions is hard?

- If content is unavailable, is it a proxy failure or censorship?
- If content is different, is it because
  - the censor is showing a message to justify its act?
  - the website content happens to be different for different countries?
  - our threshold for inferring text is not correct?
- Censoring devices might be stateful
Current experiments

- Focus on China and Iran
- 20 free proxies used in each country
- Four sets of measurements, each set visits a website twice within a few hours apart
- Known blocked sites + top 100 websites per Google in categories: overall, vpn, open proxy, democracy, entertainment, news, Buddhism, Christianity, Islam, Judaism, Taiwan, Tibet
Preliminary results

- Results across multiple measurements are not identical
- Different errors signify censorship for different proxies
  - Connection reset by peer, timeout, various HTTP errors including 404 forbidden
Preliminary results

- All ISPs in Iran either filter a website or not but results vary significantly across Chinese ISPs
- Some Chinese ISPs more permissive than others
- 404 forbidden and timeout most common techniques in Iran
- Connection reset by peer and 404 forbidden most common in China
- Google’s websites are now not blocked in China (ex: blogger.com)
## Blocked categories

<table>
<thead>
<tr>
<th>category</th>
<th>forbidden China</th>
<th>success China</th>
<th>forbidden Iran</th>
<th>success Iran</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democracy</td>
<td>0.424</td>
<td>0.505</td>
<td>0.007</td>
<td>0.640</td>
</tr>
<tr>
<td>Entertainment</td>
<td>0.305</td>
<td>0.624</td>
<td>0.423</td>
<td>0.319</td>
</tr>
<tr>
<td>News</td>
<td>0.296</td>
<td>0.605</td>
<td>0.328</td>
<td>0.405</td>
</tr>
<tr>
<td>Religion &gt; Buddhism</td>
<td>0.100</td>
<td>0.688</td>
<td>0.163</td>
<td>0.455</td>
</tr>
<tr>
<td>Religion &gt; Christian</td>
<td>0.103</td>
<td>0.753</td>
<td>0.166</td>
<td>0.595</td>
</tr>
<tr>
<td>Religion &gt; Islam</td>
<td>0.105</td>
<td>0.825</td>
<td>0.158</td>
<td>0.603</td>
</tr>
<tr>
<td>Religion &gt; Judaism</td>
<td>0.565</td>
<td>0.413</td>
<td>0.000</td>
<td>0.697</td>
</tr>
<tr>
<td>Taiwan</td>
<td>0.478</td>
<td>0.392</td>
<td>0.006</td>
<td>0.740</td>
</tr>
<tr>
<td>Tibet</td>
<td>0.861</td>
<td>0.084</td>
<td>0.048</td>
<td>0.675</td>
</tr>
<tr>
<td>Top 100</td>
<td>0.442</td>
<td>0.454</td>
<td>0.372</td>
<td>0.368</td>
</tr>
<tr>
<td>open proxy</td>
<td>0.363</td>
<td>0.508</td>
<td>0.371</td>
<td>0.407</td>
</tr>
<tr>
<td>vpn</td>
<td>0.400</td>
<td>0.460</td>
<td>0.347</td>
<td>0.384</td>
</tr>
<tr>
<td></td>
<td>0.370</td>
<td>0.526</td>
<td>0.206</td>
<td>0.524</td>
</tr>
</tbody>
</table>
Preliminary results

- What is blocked changes over time, supporting the need for ongoing measurements
- Some sites are blocked throughout our measurement period of about a month
- Using IPs for blocked host names was successful for a few Chinese ISPs (not tested for Iran)
Next steps and issues

- Test blocking by keywords, ports, protocols
- Test if DNS poisoning is in use
- Determine statefulness of censoring devices
- Examine censorship from the outside as in Mao’s work
- Understand mechanics and location of censoring devices