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

	China		Iran	
category	**forbidden**	**success**	**forbidden**	**success**
Democracy	0.424	0.505	0.087	0.640
Entertainment	0.305	0.624	0.423	0.319
News	0.296	0.605	0.328	0.405
Religion > Buddhism	0.100	0.688	0.163	0.455
Religion > Christian	0.103	0.753	0.166	0.595
Religion > Islam	0.105	0.825	0.158	0.603
Religion > Judaism	0.565	0.413	0.000	0.697
Taiwan	0.478	0.392	0.006	0.740
Tibet	0.861	0.084	0.048	0.675
Top 100	0.442	0.454	0.372	0.368
open proxy	0.363	0.508	0.371	0.407
vpn	0.400	0.460	0.347	0.384
	0.370	0.526	0.206	0.524

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