Which Websites are Censored, Anyway?

Zachary Weinberg

zackw@cmu.edu

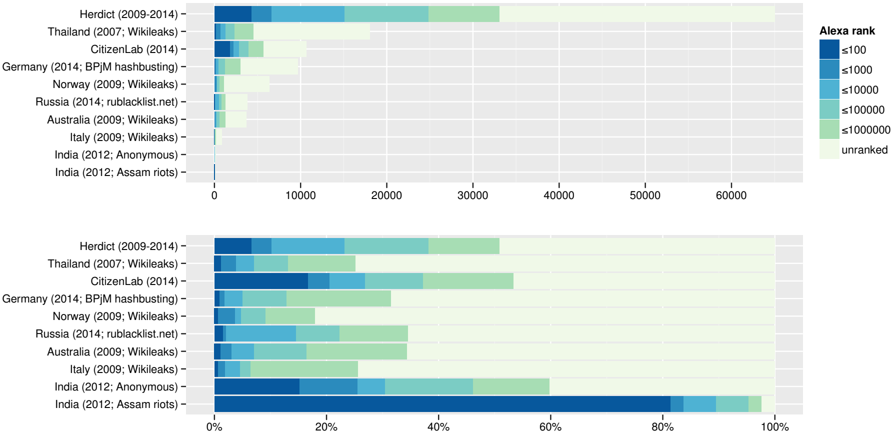
Carnegie Mellon University

17 July 2014

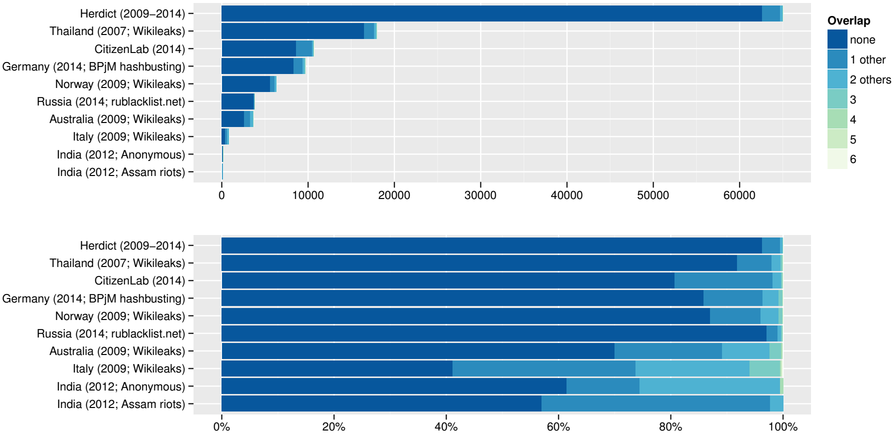
Sources

- *Herdict*, 2009–2014: 65,000 URLs reported as inaccessible (not necessarily censored) by users worldwide
- *Citizen Lab*, 2014: 10,600 URLs in synthetic list intended to probe for censorship; 2014
- *Per-country lists*, 2009, 2012, 2014: eight leaked or reverse-engineered lists of URLs and/or sites allegedly censored in those countries. Collated by `bpjmleak.neocities.org`

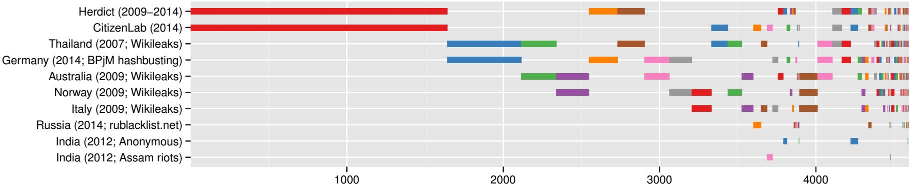
Mostly *not* the usual suspects



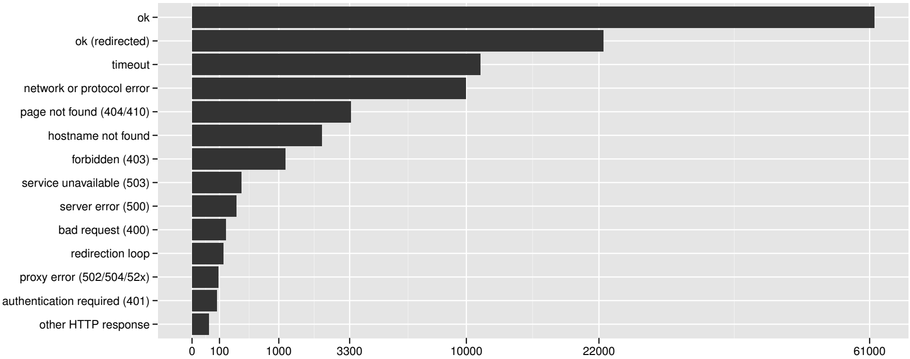
Little overlap between countries



Overlaps in more detail



Many sites no longer exist



What is being censored?

- Citizen Lab's list has categories
- The other lists do not
- Context may provide clues
- Tedious, unpleasant manual work in general
- Need to understand language of site; machine translation probably not good enough

- Please talk to me if you have ideas for automating this

Free expression and media freedom	26.41%
Human rights	6.31%
Political reform	6.06%
Anonymizers and circumvention	4.64%
Religious conversion, commentary and criticism	4.40%
Political transformation	3.71%
QUILTBAG advocacy and education	3.22%
Minority rights and ethnic content	2.97%
Pornography	2.91%
Blogging domains and blogging services	2.75%
Dating	2.66%
E-commerce	2.41%
Foreign relations and military	2.25%
Miscellaneous	2.17%
Militants, extremists and separatists	2.03%
Women's rights	1.90%
Minority faiths	1.84%
Web hosting sites and portals	1.77%
Groups and social networking	1.72%
History, arts and literature	1.59%
Multimedia sharing	1.52%
Environment	1.38%
Search engines	1.37%
Sex education and family planning	1.28%
Alcohol and drugs	1.19%
Gambling	1.06%
Peer-to-peer computing	1.06%
(14 more categories)	7.42%