Challenges in Monitoring for Web Censorship

Zachary Weinberg
zackw@cmu.edu

Carnegie Mellon University

28 July 2014

ALLOW ME TO INTRODUCING MYSELF

Zachary Weinberg
("Zack," except on paper)

network security ♦ software archaeology

email zackw@panix.com
web https://www.owlfolio.org/
pgp 82F8 54F3 CE73 174B 8B63
1740 91FC C32B 6769 AA64

- ► Born 1978—older than the average grad student
- ► Free software hacker since 1996
- ► Worked in the computer industry 1999–2005 and 2008–2010
  - ► (time off to do a master's degree in cognitive science)
- ► Got into security ca 2009, working at Mozilla
- ► Got into countersurveillance ca 2011, internship at SRI
- ► Currently at CMU, working with Nicolas Christin

My research

- Information leakage from circumvention tools (e.g. Tor)
- $\hookrightarrow$ Need to know what circumvention tools are used for
- $\hookrightarrow$ Specifically, need to know which websites are likely to be censored

Machinery I've built
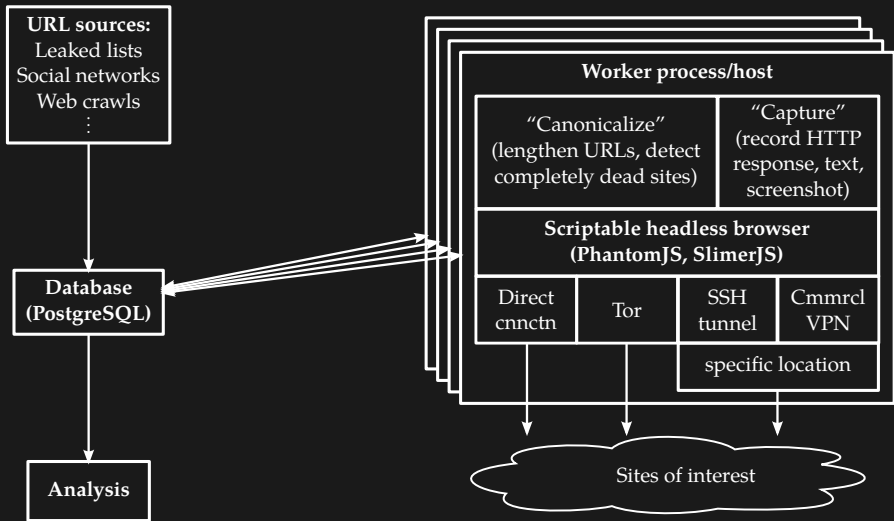
Lists found in the wild...

- *Citizen Lab*, 2014: 10,600 URLs in synthetic list intended to probe for censorship
- *Herdict*, 2009–2014: 65,000 URLs reported as inaccessible (not necessarily censored) by users worldwide
- *Per-country lists*, 2009, 2012, 2014: eight leaked or reverse-engineered lists of URLs and/or sites allegedly censored in specific countries; 100–20,000 entries each. Collated by `bpjmleak.neocities.org`
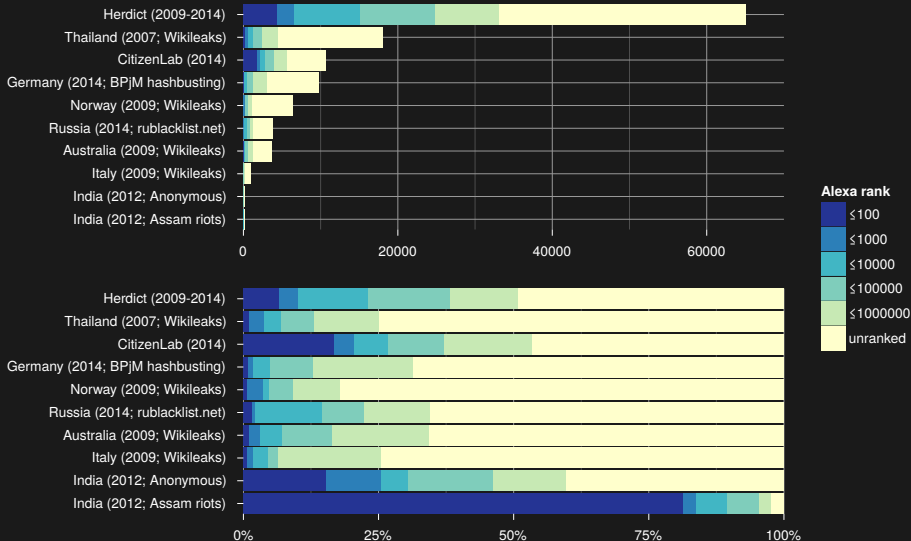
What is being censored?

- 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

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%

China vs. Hong Kong by Citizen Lab category

Categories (top to bottom):
- Free email
- Commercial sites
- Provocative attire
- Hacking
- Economic development
- Public health
- Hate speech
- Online gaming
- Alcohol and drugs
- Gambling
- Pornography
- Web hosting services
- Women's rights
- Militants, extremists and separatists
- Sex education and family planning
- Miscellaneous
- Minority rights and ethnic content
- QUILTBAG advocacy and education
- Foreign relations and military
- Voice over Internet Protocol (VoIP)
- Minority faiths
- Political reform
- Groups & social networking
- Translation
- Peer-to-peer file sharing
- Online dating
- Search engines
- Freedom of expression and media freedom
- Political transformation
- Human rights
- Environment
- History, arts and literature
- Religious conversion, commentary and criticism
- Multimedia sharing
- Anonymizers and censorship circumvention
- Blogging services

Legend:
- same result
- CN-only network error
- CN-only timeout
- HK-only network error
- HK-only timeout
- different HTTP responses

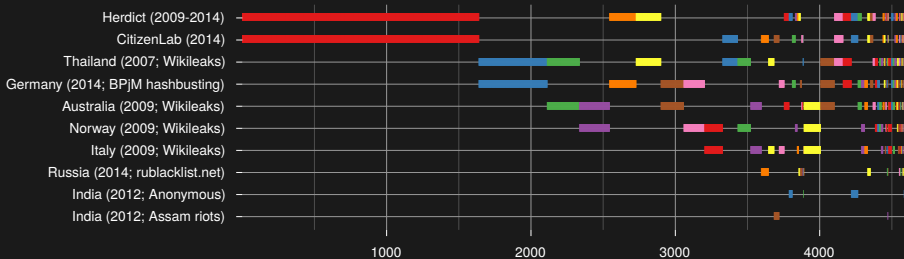x-axis: 0%, 25%, 50%, 75%, 100%

SITES ON THESE LISTS ARE IN THE "LONG TAIL"

There's not much overlap between lists
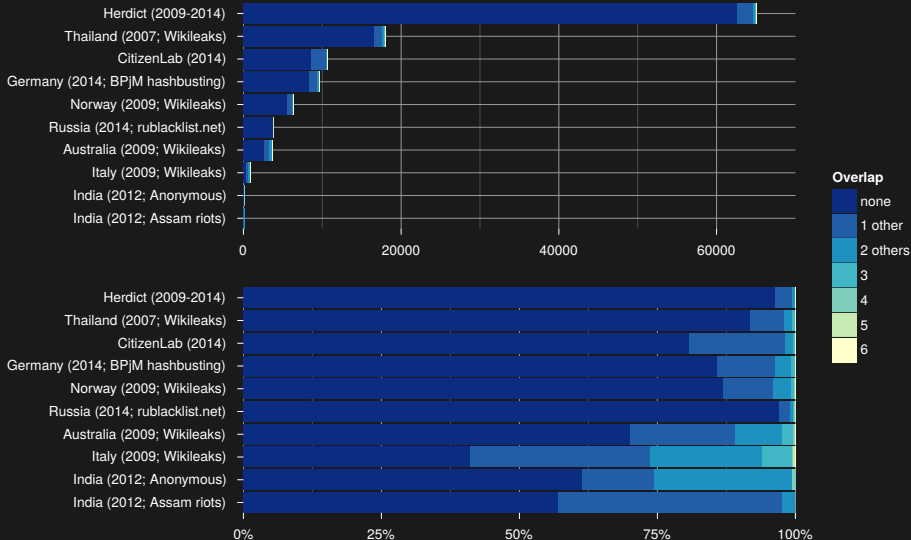
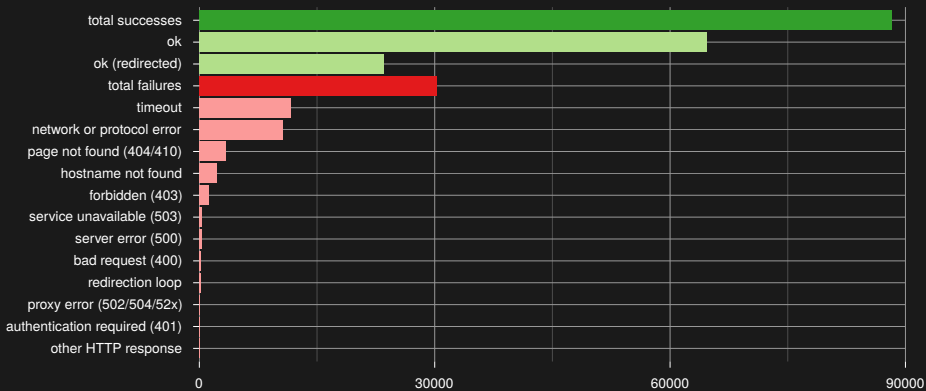# OVERLAPS IN MORE DETAIL

The largest overlap between 2 lists has 1643 entries
                                   3   116
                                   4   19
                                   5   3
                                   6   1
                                   7   1

Many sites and pages no longer exist

Accesses from the USA, June–July 2014

"Total failures" is an underestimate.

Technical challenges

To know if a page still exists and is accessible:

- ▶ You must use a feature-complete web browser.
  - ▶ Sites often try to detect and block crawlers.
  - ▶ `<meta refresh>` and JavaScript redirection are remarkably popular.
  - ▶ The "interesting" content may be entirely within `<iframe>`s.

- ▶ You may have to have a human look at it.
  - ▶ Some short-URL services, like `v.gd`, don't redirect automatically.
  - ▶ Some sites (including e.g. Youtube) respond to requests for removed content with a 200 OK.

- ▶ You may need to know what *used* to be there.
  - ▶ "according to archive.org the domain `facegoo.com` is since at least 3 years not an porn website anymore. Now it is the website of an iPhone App for fun picture manipulation." — bpjmleak
  - ▶ "Parked" domains often respond to any URL with a successful page load that has nothing to do with the original content.

Linguistic and cultural challenges

This is Dungan. It's not a Slavic language; it's a close relative of *Mandarin*, spoken by an ethnic minority living mostly in Kazakhstan these days.
It's written using Cyrillic instead of Hanzi because of the Soviet Union's literacy programs in the 1950s.

ДУНГАНСКИЙ

Чили фынли. Тянчи бянли. Йӱдянзы декэли. Йӱнцэ ба тян зо нянли. Йржынжыр залуй е щёнкэли. Быййу вон ха дони. Дихади фи чи пордини. Чучузынцди йуфи манли. Хийун гуэли, йу е жухали. Йисысыр тянчи кэ чинли. Жэту сэди хунлонлорди. Йу щягуэ, чисый жинжирдихали. Тандони за жыму чинщю, за жыму салуэ! Чёчёр гощиниди луан ду жёхуандини. Да дунбонгы гон е чӱлэли...

Йихуй йигы нэлойин мынмынди фиди дё живазылэли. Нэлойин ба живазы е мэ дешон, жё гунжи ба нэлойинди йи гӱзы мо долёли.

This image was (allegedly) censored in China a couple months ago.

weibo.com/barkcheekmark

Psychological challenges

How do *you* feel about loading up these URLs
in your web browser?

WHAT HAS BEEN SEEN

Cannot be un-seen...

- ▶ http://www.xxxslutstube.com/
- ▶ http://homemadegroupsex.com/

- ▶ http://www.incest-archive.com/
- ▶ http://www.free-horse-movies.com/

- ▶ http://www.npd-fraktion-sachsen.de/
- ▶ http://lunzenau.nw.am/
- ▶ http://www.youtube.com/user/RacialPurity666

- ▶ http://suicide-diaries.net/
- ▶ http://www.ingeb.org/
- ▶ http://www.gogru.in/
- ▶ http://e621.net/post/show/14085

Psychological challenges

How do *you* feel about loading up these URLs in your web browser?

- ► http://www.xxxslutstube.com/
- ► http://homemadegroupsex.com/

  "Honest" pornography

- ► http://www.incest-archive.com/
- ► http://www.free-horse-movies.com/

  Ickier pornography

- ► http://www.npd-fraktion-sachsen.de/
- ► http://lunzenau.nw.am/
- ► http://www.youtube.com/user/RacialPurity666

  Far-right German political advocacy?

- ► http://suicide-diaries.net/

  Instructions for suicide, in Russian

- ► http://www.ingeb.org/

  German folk songs

- ► http://www.gogru.in/

  Indian dating site

- ► http://e621.net/post/show/14085

  Cartoon/furry porn

PROBLEMS WITH THE LISTS THEMSELVES

Sample typos, technical errors, and ambiguities:

`http://hot-soccer-moms-info`	should be `.info`
`painGate.com`	fails to match `paingate.com`
`irgend.ein.name.homo.com`	"irgend.ein.name" does not act as a wildcard
`http://www.youtube.com/watch?v=PkiDwUwowrI&feature=related`	Many variations on this URL show the same video.
	In the original, the last character of "PkiDwUwowrI" was ambiguous

Unanswered questions

- How do we improve our lists of sites & pages to test?
  - Human curation, mechanical collection, and leaked filter configurations all have downsides
  - Human analysis is necessary but difficult, demoralizing, and expensive

- To what extent should we try to discern the *intent* of the censor, versus what the censor actually does?

- How can we ensure that our tests reflect reality "on the ground"?
- How can we minimize risk to research collaborators "on the ground"?