Zachary Weinberg*, Mahmood Sharif, Janos Szurdi, and Nicolas Christin

Topics of Controversy: An Empirical Analysis of Web Censorship Lists

Abstract: Studies of Internet censorship rely on an experimental technique called *probing*. From a client within each country under investigation, the experimenter attempts to access network resources that are suspected to be censored, and records what happens. The set of resources to be probed is a crucial, but often neglected, element of the experimental design.

We analyze the content and longevity of 758,191 webpages drawn from 22 different probe lists, of which 15 are alleged to be actual blacklists of censored webpages in particular countries, three were compiled using *a priori* criteria for selecting pages with an elevated chance of being censored, and four are controls. We find that the lists have very little overlap in terms of specific pages. Mechanically assigning a topic to each page, however, reveals common themes, and suggests that hand-curated probe lists may be neglecting certain frequently-censored topics. We also find that pages on controversial topics tend to have much shorter lifetimes than pages on uncontroversial topics. Hence, probe lists need to be continuously updated to be useful.

To carry out this analysis, we have developed automated infrastructure for collecting snapshots of webpages, weeding out irrelevant material (e.g. site "boilerplate" and parked domains), translating text, assigning topics, and detecting topic changes. The system scales to hundreds of thousands of pages collected.

Received 2016-05-31; revised 2016-08-01; accepted 2016-08-01.

1 Introduction

The earliest serious attempts by national governments to censor the Internet date to the early 1990s, just as the Internet was evolving into a widely-available medium of mass communication [55]. Despite concerted efforts to defend free expression, online censorship has become more and more common. Nowadays, firewalls and routers with "filtering" features are commonplace [27], and these are applied by governments worldwide to prevent access to material they find objectionable [1, 21].

Without access to inside information, determining *what* is censored is a research question in itself, as well as a prerequisite for other investigations. Official policy statements reveal only broad categories of objectionable material: "blasphemy," "obscenity," "lèse-majesté," "sedition," etc. [21, 59]. The exact list of blocked sites, keywords, etc. is kept secret, can be changed without notice, and may deviate from the official policy [28].

The experimental techniques for determining what is censored are all variations on *probing*: attempting to access network resources that are suspected to be censored from a client within the country under investigation, and recording what happens. The *probe list*, the set of resources to be probed, is a crucial element of the experimental design. Previous studies have based their lists on a variety of data sources, such as manually-selected material known to be politically sensitive in particular countries [61], crowdsourced reports of inaccessible sites [8], "leaks" from government agencies [15], and data extracted from deployed "filtering" software [38].

The central focus of this paper is evaluating the quality of existing probe lists, and determining how they can be improved. We propose the following five criteria for a high-quality probe list:

Breadth A good list includes many different types of potentially-censored material. Hand-compiled probe lists reflect the developers' interests, so they may over-investigate some types of potentially censored material and under-investigate others. Deviations from a country's official policy will only be discovered by a probe list that is not limited to the official policy.

Depth A good list includes many sites for each type of material, so that it will reveal how thoroughly that material is censored in each target country, and the boundaries of the category. This is especially important when one list is to be used to probe the policies of many different countries, because even when two countries declare the same official policy, the actual set of sites blocked in each country may be different.

***Corresponding Author: Zachary Weinberg:** Carnegie Mellon University, zackw@cmu.edu
Mahmood Sharif: Carnegie Mellon University, mahmoods@cmu.edu
Janos Szurdi: Carnegie Mellon University, jszurdi@cmu.edu
Nicolas Christin: Carnegie Mellon University, nicolasc@cmu.edu

Freshness A good list includes sites that are currently active, and avoids those that are abandoned. Sophisticated censors devote more effort to recently published content. China's "Great Firewall," for instance, not only adds sites to its blacklist within hours of their becoming newsworthy, but drops them again just as quickly when they stop being a focus of public attention [1, 19, 73]. Thus, an outdated probe list would underestimate the effectiveness of censorship in China.

Conversely, less sophisticated censors may be content to use off-the-shelf, rarely-updated blacklists of porn, gambling, etc. sites, perhaps with local additions. A recent crackdown on pornography in Pakistan led to a 50% reduction in consumption, but the remaining 50% simply shifted to sites that had escaped the initial sweep—and the censors did not update their blacklist to match [44]. Thus, an outdated probe list would overestimate the effectiveness of censorship in Pakistan.

Efficiency A good list can be probed in a short time, even over a slow, unreliable network connection. This is most important when attempting to conduct fine-grained measurements, but a list that is too large or bandwidth-intensive might not be usable at all.

Efficiency, unfortunately, is in direct tension with breadth and depth: the easiest way to make a probe list more efficient is to remove things from it. As we discuss in Section 4.1, efficiency also suffers if one seeks to collect more detailed information from each probed site.

Ease of maintenance A good list requires little or no manual adjustment on an ongoing basis. This is obviously in tension with freshness, and provides a second reason to keep lists short. The OpenNet Initiative's complete probe list contains roughly 12,000 URLs, and has only been updated a handful of times since 2014.

1.1 Topic analysis: a way forward

Mechanical analysis of the topics of Web pages that are censored, or suspected to be censored, can assist both in improving probe lists to satisfy all five of the above criteria, and in interpreting the results of probes.

Normally, censors' reasons for blocking pages will have to do with their topics. As long as this is the case, topic analysis offers an explanation of why any given page is censored. It also addresses the problem of cross-country comparison: when two countries have similar policies, blocked pages from both should exhibit strongly overlapping topic sets, even if the pages themselves overlap only a little. When the reasons are not topic-based (for instance, Syria is reported to block every website with an .il (Israel) domain name, regardless of its content [15]), topic analysis cannot provide an *explanation*, but it can still detect the phenomenon: if the blocked pages from some location do not cluster in a small number of topics, then one can manually inspect them to discover what else they have in common.

Topic analysis also provides a straightforward way to keep probe lists up-to-date. Abandoned sites can be discovered by comparing the historical topic of a page with its current topic. New pages can be discovered by identifying keywords that are relevant to sensitive topics, then searching the Web for new material.

Finally, topic analysis can reveal whether a probe list is over- or under-weighting a topic. Because of the sheer size of the Web, censors will only ever manage to block a subset of the pages on any given topic. We can estimate the number of pages on a topic that will be discovered by censors as a function of the popularity of that topic in the Web at large, the sensitivity of the topic, and the typical lifespan of pages on that topic. Probe lists can then include just as many pages as are necessary for reliable detection, and no more.

1.2 Contributions

In this paper, we analyze the uncensored contemporary and historical topics of the pages included in 22 lists, containing both sensitive and control material, as described in Section 3. We model the survival chances of all pages as a function of their topic and age. Using this information, we highlight where curated list development may have missed something important, and we discuss ways it can be done better in the future.

In support of our research, we develop a number of refinements to the standard techniques for capturing a "snapshot" of a web page and determining its contents. We use a full-featured headless web browser to ensure fidelity to what a human would see, we filter out navigation boilerplate, advertising, and the like, and we can reliably detect parked domains.

The remainder of the paper proceeds as follows. We survey previous efforts in Section 2, and describe the lists we are evaluating in Section 3. Sections 4, 5, and 6 present our methodology for data collection, preprocessing, and analysis. Our results appear in Section 7, and our conclusions in Section 8.

2 Previous Work

The earliest academic study of Internet censorship we are aware of is a 2002 case study of porn-blocking "filters" used in schools and libraries in the USA [50]. Its authors were concerned that these filters might misclassify sexual health information as pornographic. To check, they manually compiled a set of health- and porn-related keywords, expanded it to a list of 4,000 URLs by searching the Web without any filters active (2,500 URLs were health-related, 500 pornographic, and 1,000 neither) and then attempted to visit all of the pages with filters in place.

This methodology—probing the behavior of a filter with keyword searches, specific URLs known to contain sensitive content, or both—is still standard for censorship case studies. China has received the most attention [4, 18, 47, 67, 70]. Similar studies have been published for Iran [6], Pakistan [44, 45], and Turkey [45]. The OpenNet Initiative (ONI) regularly surveys roughly 80 countries worldwide [21, 22, 59].

The same methodology also underlies broader studies. One line of research investigates inter-country variation in the censorship *mechanism*: for instance, whether censorship mainly interferes with DNS lookups or subsequent TCP connections, and whether the end-user is informed of censorship [63]. In some cases, it has been possible to identify the specific "filter" in use [20, 34]. Another line aims to understand what is censored and why [1], how that changes over time [3, 28], how the degree of censorship might vary within a country [68], and how people react to censorship [37, 38].

Despite the central position of keyword and URL probe lists in all of these studies, relatively little attention has been paid to the contents of the lists, or how they are developed. Far more effort has gone into refining the methodology of the probes themselves [14, 20, 24, 26, 34, 52, 63, 68]. A few studies have dug into "leaks" of inside information, which allow researchers to see what actually *is* censored and perhaps why. Recently this has occurred for backbone filters in Syria [15] and Pakistan [37], and the TOM-Skype chat client's internal keyword filter [38]. However, all of these studies still took the list as a means to an end, not an object of research in itself.

One notable exception, and the effort closest to our own research, is the ConceptDoppler project [19]. The authors attempted to refine a keyword list, using latent semantic analysis to generate new keywords from a "seed set" of known-filtered keywords. Their goal was to develop a system that could track the time evolution of a keyword blacklist in real time, so that it could be correlated with news events. While their initial results were promising, the algorithm required a great deal of manual validation and adjustment, and we are not aware of any follow-up work.

3 Lists tested

We studied 758,191 unique URLs drawn from 22 test lists (shown in Table 1). Only one was created to be used as a probe list [61], but another 15 are (allegedly) actual blacklists used in specific countries, and two more have algorithmic selection criteria that should be positively correlated with censorship. The remaining four are control groups. One should be *negatively* correlated with censorship, and the others are neutral.

Due to the sheer size and diversity of the global Web, and the large number of pages that are not discoverable by traversing the link graph [9], any sample will inevitably miss something. We cannot hope to avoid this problem, but drawing our sample from a wide variety of sources with diverse selection criteria should mitigate it.

3.1 Potentially censored

Pages from these lists should be more likely than average to be censored somewhere.

Blacklists and pinklists These documents purport to be (part of) actual lists of censored URLs in some countries. Most are one-time snapshots; some are continuously updated. They must be interpreted cautiously. For instance, the leaked "BlueCoat" logs for Syria [15] list only URLs that someone tried to load; there is no way of knowing whether other URLs are also blocked, and one must guess whether entire sites are blocked or just specific pages.

This study includes 15 lists from 12 countries, for a total of 331,362 URLs. Eight of them include overwhelmingly more pornography than anything else; we will refer to these as *pinklists* below. (All eight do include some non-pornographic sites, even though six of them are from countries where the ostensible official policy is *only* to block pornography.) The other seven do not have this emphasis, and we will refer to them as *blacklists* below.

OpenNet Initiative ONI is an international research institute devoted to the study of Internet censorship and

surveillance. They publish a hand-curated probe list of 12,107 URLs discussing sensitive topics [61]. The principle is that these are more likely to be censored than average, not that they necessarily *are* censored somewhere. We take this list as representative of the probe lists used by researchers in this field. 1,227 of the URLs are labeled as globally relevant, the rest as relevant to one or more specific countries.

Hand-curated lists will inevitably reflect the concerns of their compilers. The ONI list, for instance, has more "freedom of expression and media freedom" sites on the list than anything else.

Herdict Herdict [8] is a service which aggregates worldwide reports that a website is inaccessible. A list of all the URLs ever reported can be downloaded from a central server; this comes to 76,935 URLs in total. The browser extension for making reports is marketed as a censorship-reporting system, but they do not filter out other kinds of site outage. This list includes a great deal of junk, such as hundreds of URLs referring to specific IP addresses that serve Google's front page.

Controversial Wikipedia articles and their references Yasseri et al. [72] observe that controversy on Wikipedia can be mechanically detected by analyzing the revision history of each article. Specifically, if an article's history includes many "mutual reverts," where pairs of editors each roll back the other's work, then the article is probably controversial. (This is a conservative measure; as they point out, Wikipedia's edit wars can be much more subtle.) They published lists of controversial Wikipedia articles in 13 languages. We augmented their lists with the external links from each article. This came to a total of 105,181 URLs.

3.2 Controls

These lists were selected to reflect the Web at large.

Pinboard We expect pages on this list to be *less* likely to be censored than average. It is a personal bookmark list with 3,276 URLs, consisting mostly of articles on graphic design, Web design, and general computer programming, with the occasional online shopfront.

Alexa 25K Alexa Inc. claims that these are the 25,019 most popular websites worldwide; their methodology is opaque, and we suspect it over-weights the WEIRD (Western, Educated, Industrialized, Rich, and Democratic [31]) population. Sensitive content is often only of interest to a narrow audience, and the popularity of major global brands gives them some protection from censorship, so sites on this list may also be less likely to be censored than average.

Twitter Another angle on popularity, we use a small (less than 0.1%) sample of all the URLs shared on Twitter from March 17 through 24, 2014, comprising 30,487 URLs shared by 27,731 user accounts. Twitter was chosen over other social networks because, at the time of the sample, political advocacy and organization via Twitter was fashionable.

Common Crawl Finally, this is the closest available approximation to an unbiased sample of the entire Web. The Common Crawl Foundation continuously operates a large-scale Web crawl and publishes the results [58]. Each crawl contains at least a billion pages. We sampled 177,109 pages from the September 2015 crawl uniformly at random.

3.3 Overlap Between Lists

We begin our investigation by comparing the probe lists to each other, using the Jaccard index of similarity: $J = \frac{|A \cap B|}{|A \cup B|}$ for any two sets A and B. It ranges from 0 (no overlap at all) to 1 (complete overlap).

Table 1 shows the Jaccard indices for each pair of lists, comparing full URLs. It is evident that, although there is some overlap (especially among the pinklists, in the upper left-hand corner), very few full URLs appear in more than one list. There is more commonality if we look only at the hostnames, as shown in Table 2. The pinklists continue clearly to be more similar to each other than to anything else. The blacklists, interestingly, continue not to have much in common with each other. And, equally interestingly, all the other lists—regardless of sampling criteria—have more in common with each other than they do with most of the blacklists and pinklists. This already suggests that manually curated lists such as ONI's may not be digging deeply enough into the "long tail" of special-interest websites.

While we can see that there are patterns of similarities, Tables 1 and 2 do not reveal *what* some lists have in common with each other. To discover that, we must study the content of each page, which is the task of the rest of this paper.

Table 1. Jaccard coefficients for list similarity, by URL

		(size)	aus	dnk	fin	deu	ita	nor	th1	tur	in1	in2	rus	syr	th2	th3	gbr	oni	hdk	wki	pin	alx	twi	ccr
pinklist	Australia 2009	5 130	1	<	.01	.02	.05	.03	.01	<			<	<			<	<	<			<	≪	
	Denmark 2008	7 402	<	1	.08	<	<	.12	<	<			≪				≪		≪					
	Finland 2009	1 336	.01	.08	1	<	.03	.04	<	<			<				≪		<					
	Germany 2014	13 174	.02	<	<	1	<	<	.02	.01			<				<	<	<	≪		<	≪	
	Italy 2009	1 078	.05	<	.03	<	1	.03	<	<			<				<	<	<			<		
	Norway 2009	14 022	.03	.12	.04	<	.03	1	<	<			<				<		<					
	Thailand 2007	26 789	.01	<	<	.02	<	<	1	.01			<	<			<	<	<	≪		<	≪	
	Turkey 2015	172 971	<	<	<	.01	<	<	.01	1	≪		<	≪			.02	<	<	≪	≪	<	<	≪
blacklist	India 2012 (Anonymous)	214								≪	1						<	<	<					
	India 2012 (Assam riots)	103										1						<						
	Russia 2014	4 482	<	≪	<	<	<	<	<	<			1				<	<	<			<		
	Syria 2015	12 428	<						<	≪				1			<	<	<	<	≪	<	≪	
	Thailand 2008	1 298													1	<			<					
	Thailand 2009	408													<	1								
	Great Britain 2015	87 032	<	≪	≪	<	<	<	<	.02	<		<	<			1	<	.03	<	<	<	<	≪
probe list	OpenNet Initiative 2014	12 107	<			<	<		<	<	<	<	<	<			<	1	.02	<	<	.01	<	≪
crowdsourced	Herdict 2014	76 935	<	≪	<	<	<	<	<	<	<		<	<	<		.03	.02	1	<	<	.04	<	≪
sampled	Wikipedia controv. 2015	105 181				≪			≪	≪				<			<	<	<	1	≪	<	<	≪
neg. control	Pinboard 2014	3 876								≪				≪			<	<	<	≪	1	<	≪	≪
popular	Alexa 2014	25 019	<			<	<		<	<			<	<			<	.01	.04	<	<	1	<	≪
	Tweets 2014	40 198	≪			≪			≪	<				≪			<	<	<	<	≪	<	1	≪
generic	Common Crawl 2015	177 109								≪							≪	≪	≪	≪	≪	≪	≪	1

<: smaller than 0.01. ≪: smaller than 0.0001. Blank: zero.

Table 2. Jaccard coefficients for list similarity, by hostname

		(size)	aus	dnk	fin	deu	ita	nor	th1	tur	in1	in2	rus	syr	th2	th3	gbr	oni	hdk	wki	pin	alx	twi	ccr
pinklist	Australia 2009	1 752	1	.01	.04	.03	.08	.04	.02	<		<	<	<	<	<	<	<	<	<	<	<	<	<
	Denmark 2008	7 402	.01	1	.08	<	<	.19	<	<			<				≪		<					
	Finland 2009	1 336	.04	.08	1	<	.04	.07	<	<			<	<			≪		<					≪
	Germany 2014	6 199	.03	<	<	1	<	.01	.02	.01		<	<	<	<	<	<	<	<	<	<	<	<	<
	Italy 2009	539	.08	<	.04	<	1	.03	<	<			<				<	<	<			<		≪
	Norway 2009	7 011	.04	.19	.07	.01	.03	1	<	<			<				<		<	≪				≪
	Thailand 2007	11 880	.02	<	<	.02	<	<	1	.01		≪	<	<	<	<	<	<	<	<	<	<	<	<
	Turkey 2015	172 971	<	<	<	.01	<	<	.01	1	≪	≪	<	<	≪	≪	.02	<	<	<	≪	<	<	<
blacklist	India 2012 (Anonymous)	203								≪	1			<			<	<	<	≪	<		<	<
	India 2012 (Assam riots)	21	<			<			≪	≪		1	<	<	<	<	≪	<	<	<	<	<	<	<
	Russia 2014	1 994	<	<	<	<	<	<	<	<		<	1	<	<	<	<	<	<	<	<	<	<	<
	Syria 2015	6 526	<		<	<			<	<	<	<	<	1	<	<	<	.02	<	<	<	<	<	<
	Thailand 2008	104	<			<			<	≪		<	<	<	1	.21	≪	<	<	<	<	<	<	<
	Thailand 2009	94	<			<			<	≪		<	<	<	.21	1	≪	<	<	<	<	≪	<	≪
	Great Britain 2015	79 510	<	≪	≪	<	<	<	<	.02	<	≪	<	<	≪	≪	1	<	.03	<	<	<	<	<
probe list	OpenNet Initiative 2014	10 016	<			<	<		<	<	<	<	<	.02	<	<	<	1	.02	.02	<	.02	<	<
crowdsourced	Herdict 2014	70 528	<	<	<	<	<	<	<	<	<	<	<	<	<	<	.03	.02	1	.02	<	.04	.01	.02
sampled	Wikipedia controv. 2015	27 410	<			<		≪	<	<	≪	<	<	<	<	<	<	.02	.02	1	<	.04	.01	.03
neg. control	Pinboard 2014	2 495	<			<			<	≪	<	<	<	<	<	<	<	<	<	<	1	<	<	<
popular	Alexa 2014	24 977	<			<	<		<	<		<	<	<	<	≪	<	.02	.04	.04	<	1	.02	.05
	Tweets 2014	12 504	<			<			<	<	<	<	<	<	<	<	<	<	.01	.01	<	.02	1	<
generic	Common Crawl 2015	47 042	<		≪	<	≪	≪	<	<	<	<	<	<	<	≪	<	<	.02	.03	<	.05	<	1

<: smaller than 0.01. ≪: smaller than 0.0001. Blank: zero.

4 Data Collection

We collected both *contemporary* and *historical* snapshots of each page on our lists. As the names imply, contemporary snapshots show the page as it currently is, whereas historical snapshots look back in time. Contemporary data is sufficient to evaluate webpage availability and topic. Historical data reveals pages whose topic has changed since they were entered onto a blacklist; it helps us discover the topic of pages that no longer exist; and it allows us to compute page availability over time, as well as topic changes over time.

In this study, we are not attempting to discover whether any given page actually is censored anywhere. Rather, we seek to explain *why* a list of pages (such as the ONI probe list, or one of the blacklists) is more or less likely to be censored, by comparing its topic distribution to a reference list (such as the Common Crawl sample of the entire Web). For this purpose, we need *uncensored* snapshots of each page. Therefore, all snapshots were collected by computers in commercial data centers in the USA, where Internet censorship conflicts with the constitutional protection of freedom of speech. There have been occasional moves toward blocking access to "obscene" material in the USA [60], but we are not aware of any filtering imposed on the commercial servers we use in this study.

4.1 Contemporary Data Collection

We used an automated Web browser, PhantomJS [32] to collect contemporary snapshots of each page. PhantomJS is based on WebKit, and supports roughly the same set of features as Safari 6.0. Relative to "bleeding edge" browsers, the most significant missing features involve multimedia content (video, audio, etc.), which we would not collect anyway, for legal reasons (see below). A controller program started a new instance of PhantomJS for each page load, with all caches, cookie jars, etc. erased.

An automated browser offers major advantages over traditional "crawling" using an HTTP client that does not parse HTML or execute JavaScript. An increasing number of pages rely on JavaScript to the point where a client that does not run scripts will see none of the intended content. Also, markup ambiguities and errors are handled exactly as a human-driven browser (Safari) would. Downstream processing receives only well-structured, canonicalized HTML documents. Finally, it is harder for the server to detect that it is being accessed by an automated process, which might cause it to send back different material than a human would receive [66].

This approach also has disadvantages. The most significant is its cost in time and computer power. The data-collection host could sustain an average page-load rate of approximately 4 pages per second, with the limiting factor being PhantomJS's substantial RAM requirements. A well-tuned traditional crawler, by contrast, can sustain an average page-load rate of 2,000 pages per second with roughly equivalent hardware resources [2]. We also suffer from a much larger set of client bugs. 6,980 attempted page loads (0.92%) caused PhantomJS to crash. Finally, it is still possible for sites to distinguish PhantomJS from a "real" browser. Some sites block this kind of close mimic, while allowing obvious web crawlers access. For instance, LinkedIn blocked us from accessing user profile pages and job listings.

Our collector ignores `robots.txt`, because a human-driven browser would do the same. Instead, we avoid disruptive effects on websites by randomizing the order of page loads, so that no website sees a large number of accesses in a short time. Also, we do not traverse any outbound hyperlinks from any page, which reduces the odds of *modifying* sites by accessing them. For legal reasons, our collector does not load images and videos, nor does it record how the pages would be rendered. While HTML sources are safe, there exist images that are illegal to possess, even unintentionally, in the USA.

Ideally, contemporary data collection should occur at a single point in time, but this is impractical, given the volume of data we are acquiring. Most of our contemporary data was collected over a two-month period ranging from September 21 through December 3, 2015. For efficiency, we imported HTML directly from Common Crawl's data release rather than re-crawling each page ourselves. These pages were collected between July 28 and August 5, 2015. More importantly, Common Crawl is a traditional crawler, so these pages' contents may be less accurately recorded.

4.2 Historical Data Collection

Our historical snapshots were all collected by the Internet Archive's "Wayback Machine" [33]. The Archive began recording Web pages in 1996. They offer HTTP-based APIs for retrieving all the dates where they have a snapshot of a particular page, and then for retrieving the page as they saw it on a particular date. We used these APIs to retrieve snapshots at one-month intervals

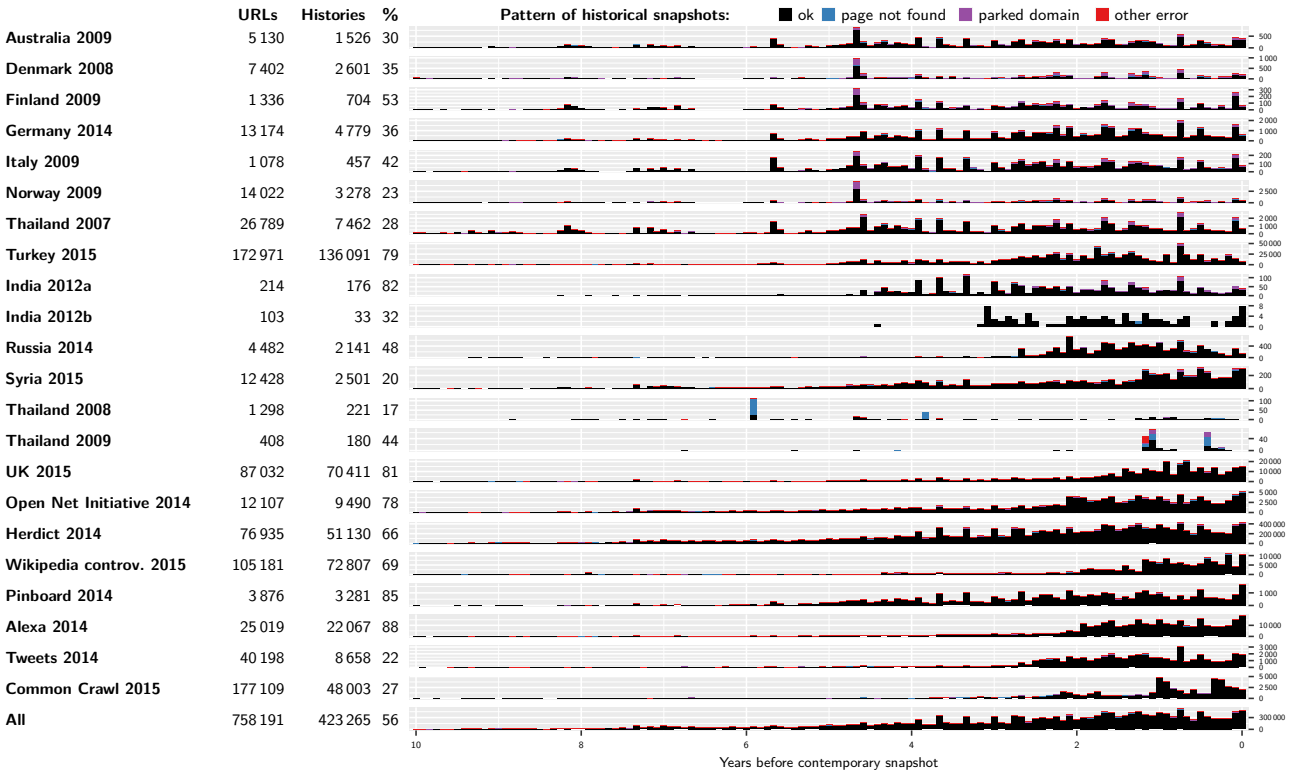

Fig. 1. Availability of historical snapshots for the pages on each source list

(whenever possible), running backward in time from the date of our contemporary snapshot to at least one year before the earliest date that the page appears in any of our lists. Like Common Crawl, the Wayback Machine uses a traditional crawler, so there is some loss of fidelity in historical snapshots.

The Wayback Machine has snapshots for 423,265 of the 758,192 pages in this study (55.8%), but these are not evenly distributed across all of the lists. Figure 1 shows how many of the pages on each list have historical snapshots, and how those are distributed over the 10 years before our contemporary data was collected. The Wayback Machine is more likely to collect popular and long-lived websites, and, unfortunately, this means it has less data for the sites on the pinklists and blacklists. As we will discuss in Section 7.2, we have enough data to predict the lifetime of a page as a function of its source category, but not individual sources, topics or languages.

"India 2012a" appears to be well-collected, but this is an artifact. That list consists mostly of YouTube videos; YouTube is extremely popular, so the Wayback Machine has good coverage of it. Most of the videos have been removed from YouTube (we suspect this is a case of "DMCA takedown abuse," in which a legal process intended to combat copyright infringement is applied to suppress controversial material [30]), but YouTube's "This video is no longer available" error message is served as a *successful* HTTP transaction (200 OK). Thus, the pages appear to have survived, when they haven't. Fortunately, there are only a few hundred pages affected by this artifact.

For 81,988 of the pages (10.8%), the Wayback Machine records at least one snapshot within 30 days of the earliest date when the page was entered onto one of our lists. It is at this time that the page's topic is most likely to be relevant to its chances of being censored.

5 Document preprocessing

Having collected pages, we wish to reason about their contents. For example, we wish to assign a topic to each page, and detect when this topic changes. With hundreds of thousands of pages collected, this process must be automated as much as possible.

The principal technique we use is Latent Dirichlet Allocation (LDA) clustering [11]. Our analysis pipeline (illustrated in Figure 2) includes several heuristic filtering steps before LDA, which remove irrelevant "boilerplate" and reduce the cost of model training. These are

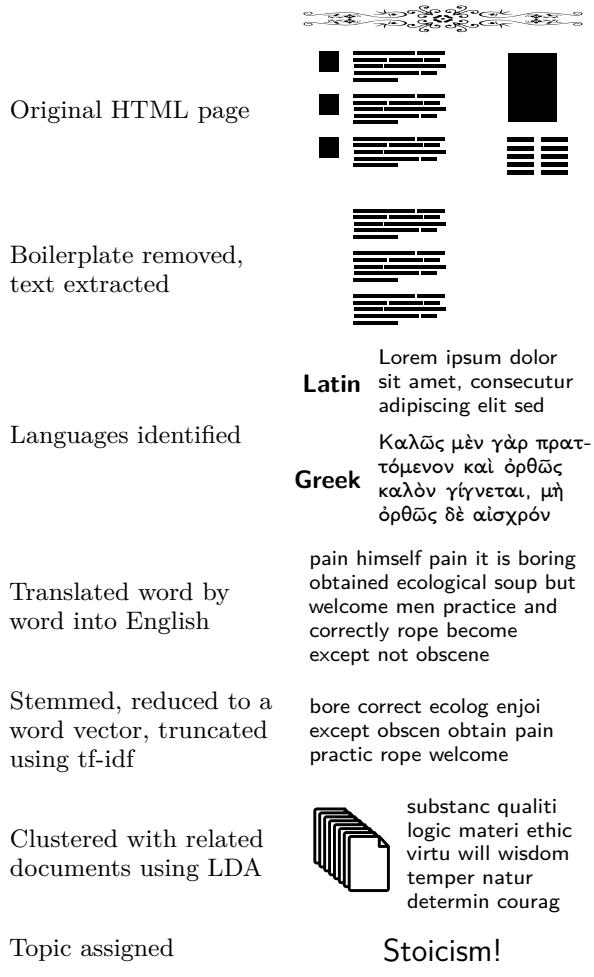

Fig. 2. Document processing pipeline

described in this section, and LDA itself is described in Section 6.

5.1 Parked Domain Detection

A *parked domain* is a placeholder website, operated by a *domain parker*, who hopes to sell the domain name eventually. It contains no meaningful content, only advertising banners and links [64]. Domain parkers often claim the names of abandoned websites, hoping to gain from visits by people looking for the former occupant. The placeholder site will bear little or no relationship to the content of the website that used to be there. However, it may parrot keywords from search queries that lead to the site. This confuses LDA, which cannot tell that the words are being repeated meaninglessly.

Therefore, we identify parked domains using a dedicated, heuristic classifier and exclude them from topic analysis. We tested two such classifiers from the literature [57, 64] and selected the one that performed best on our data; see Appendix A for details.

Contemporary data collection retrieved a non-empty, non-error HTML page for 436,832 of the 758,191 unique URLs in our full data set (57.6%). The classifier identified 30,508 (7.0%) of these pages as parked domains.

Parked domains also appear in the historical data. The Wayback Machine provided us with 2,742,788 snapshots of 423,236 URLs. Of those snapshots, 2,486,426 (90.7%) were non-empty and not error pages, and the classifier identified 54,710 (2.2%) of these as parked domains. 23,687 (5.6%) of the 423,236 URLs were parked at least once in their history.

5.2 Boilerplate Removal

Nearly all HTML documents contain "boilerplate" text which has little or nothing to do with the topic of the document itself, such as site navigation menus, copyright notices, and advertising. It may not even be in the same language as the document's main content [54]. Boilerplate varies only a little from site to site, so it can confuse semantic analysis algorithms into grouping documents that are unrelated. This problem has been recognized since 2002 [7] and the solution is to strip the boilerplate from each document prior to semantic analysis. Unfortunately, the most widely used algorithms for stripping boilerplate, such as Readability [49] and the similar "reader view" features in Chrome, Firefox, and Safari, depend on the standard semantics of HTML elements. In a large sample of not necessarily well-structured documents, this is not a safe approach. Some algorithms also make strong assumptions about the document language [25] or require several pages from the same site [54].

We developed a hybrid of the boilerplate removal algorithms described in Lin et al. [40] and Sun et al. [56]. These are completely language-neutral, use HTML element semantics only as hints, and in combination, require no manual tuning. Their basic logic is that heavily marked-up text is more likely to be boilerplate.

The hybrid algorithm merges subtrees of the parsed HTML into a tree of "blocks," each of which represents a contiguous run of text. Blocks are bigger than paragraphs but usually smaller than sections. Each block is assigned a *text density* score, which is the total number of text characters in the block, divided by the logarithm of the total number of markup characters in the

block. Stylistic markup (bold, italic, etc.) does not count, and invisible HTML elements (scripts, etc.) are completely discarded. After the entire page has been scored, the algorithm identifies the *least* dense block that contains the *most* dense block. This block's density score is the "threshold." Every block that is less dense than the threshold (that is, it contains more markup and less text) is removed from the page. Finally, all remaining markup is stripped, leaving only text.

5.3 Language Identification and Translation

LDA topic models detect semantic relationships between words based on their co-occurrence probabilities within documents. Therefore, it is necessary for all documents to be in the same language. Multi-lingual versions of LDA exist, but they are either limited to two languages [12], or they require all documents to be available in all languages, with accurate labeling [42]. Our data meets neither condition, so instead we mechanically translated as much text as possible into English.

After boilerplate removal, we used CLD2 [53] to determine the languages of each document and divide multilingual documents into runs of text in a single language. We then used Google Translate's API [29] to translate as much text as possible into English. At the time of writing, CLD2 can detect 83 languages with accuracy higher than 97%, and Google Translate can translate 103 languages into English; neither set is a superset of the other. 11.5% of all words were unrecognized or untranslatable; the bulk of these were nonwords (e.g. long strings of digits) and errors on CLD2's part. In a bilingual document, for instance, CLD2 frequently gets each split point wrong by a couple words, or tags small runs of one language as "unknown." Only 29,234 documents (0.8%) were completely untranslatable.

Google charges US$20 to translate a million characters. After boilerplate removal, the 4,355,234 unique pages in our database (including both contemporary and historical snapshots) add up to 13.3 *trillion* characters; translating each document in full would have cost $260,000, which was beyond our budget. Instead, we reduced each document to a "bag of words," and then translated each word in isolation, which cost only $3,700. This required us to "segment" text into words, which is nontrivial in languages written without spaces between the words. For Chinese we used the Stanford segmenter [17]; Japanese, MeCab [39]; Vietnamese, dongdu [5]; Thai, libthai [36]; Arabic and related languages, SNLP [43]; all others, NLTK [10].

Because our data set is so large, we needed to truncate the translated word vectors to complete training in a reasonable amount of time. After translation, we reduced all words to morphological stems using the Porter stemmer [48]. We then used *term frequency-inverse document frequency* (tf-idf, [51]) to select terms with a high degree of salience for each document, preserving terms whose combined *tf-idf* constituted (at least) 90% of the total. After pruning, the median size of a word vector was 37 words.

6 Topic Analysis

We used the MALLET implementation of LDA [41, 46, 71] to cluster documents into topics.

We used the contemporary data for training and selecting the topic models. Half of the collected documents were used for training, and the remainder were used for model-selection. We trained models with $N \in \{100, 150, \ldots, 250\}$ topics and $\alpha \in \{0.1, 0.5, 1, 5, 10, 100\}$ (α controls the sparsity of the topic assignment), and selected the max-likelihood model, following the procedure described by Wallach et al. [65]. We found the parameters $N = 100$ and $\alpha = 5$ to be optimal.

After model training, two researchers reviewed the top words associated with each topic and labeled the topics. A colleague not otherwise involved with the research scored inter-coder agreement between the labels, which came to 87%. Disagreements were resolved by discussion between the researchers.

To capture the complexities of modern web pages (e.g., dynamically updated contents, mashups, etc.), rather than assigning a single topic to each given page, we assigned it a vector of probabilities over all N topics. For instance, a news website front-page containing article snippets about sports and politics would have those topics ("sports," "news," "politics") assigned relatively high probabilities, perhaps 0.2, 0.4 and 0.35. Other topics would receive probabilities very close to zero.

LDA found several topics with identical labels. This is a known limitation of LDA when the training dataset is skewed toward certain topics. The algorithm will split those topics arbitrarily in order to make all of the clusters roughly the same size [69]. We solved this problem by manually merging topics that have similar labels and summing their probabilities. For instance, suppose that topics 26 and 61 were both labeled "news," and that

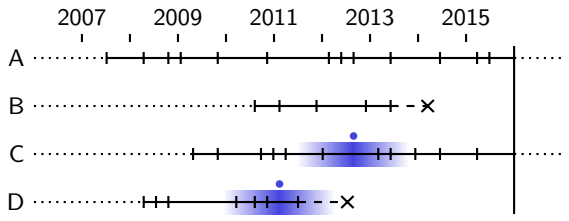

Fig. 3. The life cycles of four hypothetical websites. Tick marks are Wayback Machine snapshots; the vertical bar is our contemporary data capture; blue shaded areas indicate possible censorship events.

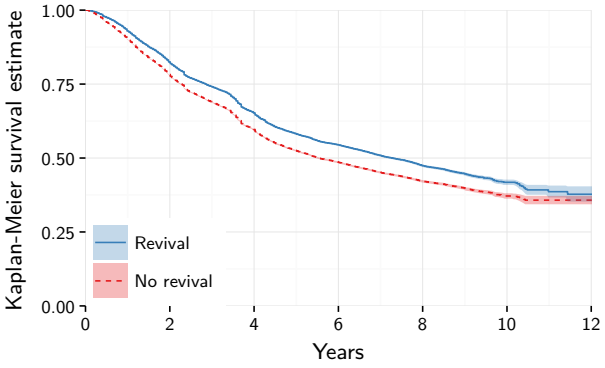

Fig. 4. Comparison of the two approaches to page revival. Shading shows confidence intervals.

a page has probability 0.24 for topic 26 and 0.56 for topic 61. These topics would be combined into a single "news" topic, and the page's weight for the combined topic would be 0.8.

Using this procedure, our initial set of 100 topics was reduced to 64 merged topics, and then further grouped into nine categories. Two artifical topics were added to account for documents that could not be processed by LDA at all. The final set of topics and categories is shown in Table 3 along with measures of the bias of lists and languages toward each topic. We discuss the topic assignments further in Section 7.

6.1 Survival Analysis

Our data on the life cycle of websites is unavoidably incomplete. Figure 3 shows four hypothetical cases which illustrate the problem. In no case do we know when a page was created, only when it first came to the attention of the Wayback Machine. If a page survives to the present (A, C), we do not know how much longer it will continue to exist. If it was abandoned (B, D), we only know that this happened within an interval between two observations. If a page appears on a censorship blacklist (C, D), we know when this happened (blue dot) but we can only guess at how long the page was censored (blue shaded area).

Survival analysis [35] is a set of statistical techniques originally developed for predicting the expected lifespan of patients with terminal illnesses. Because medical studies often suffer from exactly the same kinds of gaps in their data—one doesn't usually know how long a tumor was present before it was diagnosed, for instance—survival analysis is prepared to deal with them. However, to use these techniques we must define what it means for a page to "die." Clearly, if the site is shut down or turns into a parked domain, that should qualify. Less obviously, we also count topic change as "death." This is because, after a censored page has changed topic, it no longer provides the same kind of sensitive material that it used to, so it can no longer be considered "fresh." For example, the blog http://amazighroots.blogspot.com was taken over by spammers in 2014, and its apparent topic changed from "news and politics" to "food." Probing such pages longer reveals whether the censor cares about that kind of sensitive material. It may instead reveal how diligent the censor is about updating their blacklist, but this was not the original goal.

Unlike medical patients, webpages may be "dead" only temporarily, due to server crashes, vandalism, the owners forgetting to renew the domain registration, and so on. Stock survival analysis does not allow for this possibility. To handle it, we calculated every survival curve two ways: first, assuming that revivals never happen (once a site "dies" it is treated as staying that way, even if we have evidence to the contrary) and second, allowing sites to revive even after an arbitrarily long hiatus. The first approach gives an underestimate of survival times, the second, an overestimate. Figure 4 shows, over all pages we monitored, that on average, the differences between both approaches are relatively small; and that our error ranges are small.

6.2 Detection of Topic Changes

As discussed in Section 6.1, pages can cease to be relevant to a censorship probe list by being taken down entirely, or by changing their topic to the point where they no longer contain sensitive material. Therefore, to evaluate the freshness of a probe list we need to detect topic changes.

Existing algorithms for detecting *any* change in a webpage (e.g. [16]) are too sensitive for our purposes. Even looking for changes in the most probable topic

chosen by LDA is too sensitive. The most probable topic assigned to the front page of a news website changes several times every day, as new articles appear, but it is still the front page of a news website.

Instead, we compare the entire sets of probable topics that were assigned to a pair of observations. Specifically, if T_1 and T_2 are the topic probability vectors assigned to a pair of observations, let $S_1 = \{i : T_{1i} \geq p\}$ and $S_2 = \{i : T_{2i} \geq p\}$, that is, the respective sets of topic indices for which the assigned probabilities are greater than p. Then, the page's topic is judged to have changed if $|S_1 \cap S_2| < m$, that is, the intersection of the topic sets is smaller than m.

This algorithm has two parameters, p and m. Its performance is also affected by the LDA sparsity parameter α and the number of topics N. To tune these parameters, we used a manually chosen development set of 30 pairs of observations whose topics were the same, and another set of similar size of observation pairs whose topics were different. We found that for $p = 0.05$, $m = 1$, $\alpha = 10$, and $N = 100$ the algorithm achieved perfect accuracy on the development set. On a second randomly-chosen evaluation set, with 50 pairs of observations whose topics were the same and 50 whose topics were different, the algorithm achieved 97.8% recall and 86% precision.

7 Results

With every page assigned a topic, we can now look at how topics correlate with the lists the pages came from, and the languages they're written in. With information about when pages went offline or changed their topic, we can also predict the typical lifetime of classes of pages.

7.1 Topic Correlations

To examine the correlation of topics with source lists and languages, we apply χ^2 tests of independence to the contemporary data set. Overall tests strongly confirm the hypotheses that the distribution of documents over topics is correlated with source list and with language. Coincidentally, both tests have 1365 degrees of freedom. For topic × source, $\chi^2 = 6.45 \times 10^5$ ($p < 0.001$), for topic × language, $\chi^2 = 1.33 \times 10^6$ ($p < 0.001$). We then perform post-hoc χ^2 tests on each combination of list and topic, or language and topic, using a 2×2 contingency table of the form

$$\begin{array}{c|c} w_{gt} = \sum_{u \in g} \mathbf{T}_{ut} & w_{g\neg t} = |g| - \sum_{u \in g} \mathbf{T}_{ut} \\ \hline w_{rt} = \sum_{u \in r} \mathbf{T}_{ut} & w_{r\neg t} = |r| - \sum_{u \in r} \mathbf{T}_{ut} \end{array} \quad (1)$$

where g is the selected list or language, r is the *reference* list or language (see below), and $\mathbf{T}_{ut}$ is the probability that page u belongs to topic t. (Recall from Section 6 that each page is assigned a probability vector over all topics.) There are a total of 2,904 such combinations. After Bonferroni correction, 585 of the topic-list correlations and 580 of the topic-language correlations are significant at the usual $\alpha = 0.05$ level.

However, significant correlations might still be too small to be interesting. Rather than show significance by itself, therefore, we compute the *odds ratio* for each significant cell. This statistic can be computed directly from the 2×2 contingency table above:

$$r_{t;g,r} = \frac{w_{gt}/w_{g\neg t}}{w_{rt}/w_{r\neg t}} \quad (2)$$

It is one when there is no difference between the source and the reference, greater when the group has more pages on a topic than the reference does, and smaller when it has fewer. In Table 3, we show the odds ratio for each significant comparison. Again, this considers contemporary page contents only. Blank cells are non-significant. Shades of red indicate that the topic is positively correlated with the list or language, and shades of blue indicate that it is anti-correlated.

In the left half of Table 3, we correlate the topics with the source lists, taking Common Crawl as the reference (we believe this to be the most topic-uniform of our lists; but see Appendix B for a counterargument). When two lists have red cells for the same topic, that indicates a commonality between the lists. However, when two lists have blue cells for the same topic, that means only that neither is correlated with that topic, which does not qualify as something they have in common.

We can immediately see the same three clusters of source lists that appeared in Tables 1 and 2. The blacklists have more in common with the potentially-censored lists and the controls, but when it comes to the most politically controversial categories (news, politics, religion, etc.) they tend to be concentrated on one or two specific topics, whereas the potentially-censored lists are spread over many such topics. In some cases it is apparent that a country is censoring news related to *itself*, but not other news. The Syria 2015 list includes a surprisingly large number of software-related sites; spot checking indicates that this is due to indiscriminate blocking of websites with Israeli domain names.

Table 3. Correlation of topics with languages and source lists.

The blacklists also devote more attention to specific entertainment topics than the potentially-censored lists do. Social media in particular stands out, consistent with external evidence that this is increasingly seen as a threat by censors [23]. Blocking access to video-sharing and other entertainment sites may also be meant to suppress copyright infringement and/or support local businesses over global incumbents [37].

Pornographic topics are concentrated in the pinklists and underrepresented elsewhere. All of the pinklists have some non-pornographic pages. Some of these can be explained by poor classification of image-heavy pages, and by debatable classification of, for instance, "mail-order bride" sites. However, we do see a genuine case of political censorship under cover of porn filtering: many of the pages on the Thailand 2007 list that were filed under Japan, Vietnam, or social media are discussing Southeast Asian regional politics. This was known from previous case studies of Thailand [62] and is exactly the phenomenon we designed our system to detect.

The negative control (Pinboard) is almost perfectly anticorrelated with the blacklists and pinklists. There is some overlap on software topics. This is largely due to the negative control being strongly biased toward those topics, so any software topics at all in any of the blacklists will show as overlap. Also, software-industry-focused news sites tend to be hostile to attempts to censor the Internet.

The other controls have more in common with Common Crawl than with the blacklists or pinklists. They also have more in common with the "probably censored" lists than the blacklists or pinklists; here we see that popular pages are more likely to get cited on Wikipedia or have someone bother to report an outage to Herdict. The over-weighting of some regional news topics in this group of lists may also indicate biases in Common Crawl (see Appendix B).

In the right half of Table 3, we correlate topics with the 21 most commonly used languages in our data set (and with "other languages"), taking English (which is far and away the most common) as the reference. The same caution about paired red versus blue cells applies.

News topics for specific countries are very strongly correlated with the languages spoken in those countries, and Islam correlates with languages spoken in countries where it is the most or second-most common religion. Many of the more "commercial" topics are dominated by English; this may be an artifact of data collection from the USA, since commercial sites often change their language depending on the apparent location of the client.

The "junk" topics at the bottom of the table collect various documents that we could not interpret meaningfully. Despite our efforts to weed them out early, some error messages (perhaps served with the wrong HTTP response code, so the crawler does not detect an unsuccessful page load) and webpage boilerplate creep through. Mistranslated, unintelligible, and empty documents are self-explanatory. Finally, documents that we were unable to translate are in languages that Google Translate does not support, or (in the case of Japanese) suffered from a character encoding problem that made them *appear* untranslatable to the automation. The high concentration of error messages on the pinklists probably reflects the short lifetime of pages on these lists (see Section 7.2 for more on this); the untranslatable documents on the pinklists may be an artifact of porn sites carrying far more imagery than text; the mistranslations on the blacklists probably indicate weak support for colloquial Russian and Arabic in particular.

7.2 Survival Analysis

We modeled page survival curves using Kaplan-Meier estimators [35] with right-censoring[1] and delayed entry. When death events were only known to have occurred within some interval, we substituted the midpoint of the interval. We compared survival curves using log-normal tests and Cox proportional hazard models.

As mentioned in Section 4.2, the limited data we have from the Wayback Machine is insufficient to compute Kaplan-Meier curves for each topic, or each source list. Only 55.8% of the URLs have any historical data at all, and the median number of historical snapshots per URL is only 3, with large gaps between observations. We do have enough data to compute K-M curves for each group of source lists (Figure 6) and each category of topics (Figure 7). These larger-scale clusters correspond to the horizontal and vertical divisions of the left half of Table 3, plus an extra topic category just for HTTP errors. It should be said, however, that the large gaps mean that all these curves probably overestimate survival times.

From these curves, we can see that pages hosting sensitive material (pinklist, blacklist, and probably censored; porn, software, entertainment, video, news) are significantly shorter-lived than less sensitive webpages, with the pinklist pages faring worst. (The especially

1 Survival analysis jargon uses the word "censored" to describe a particular kind of missing data.

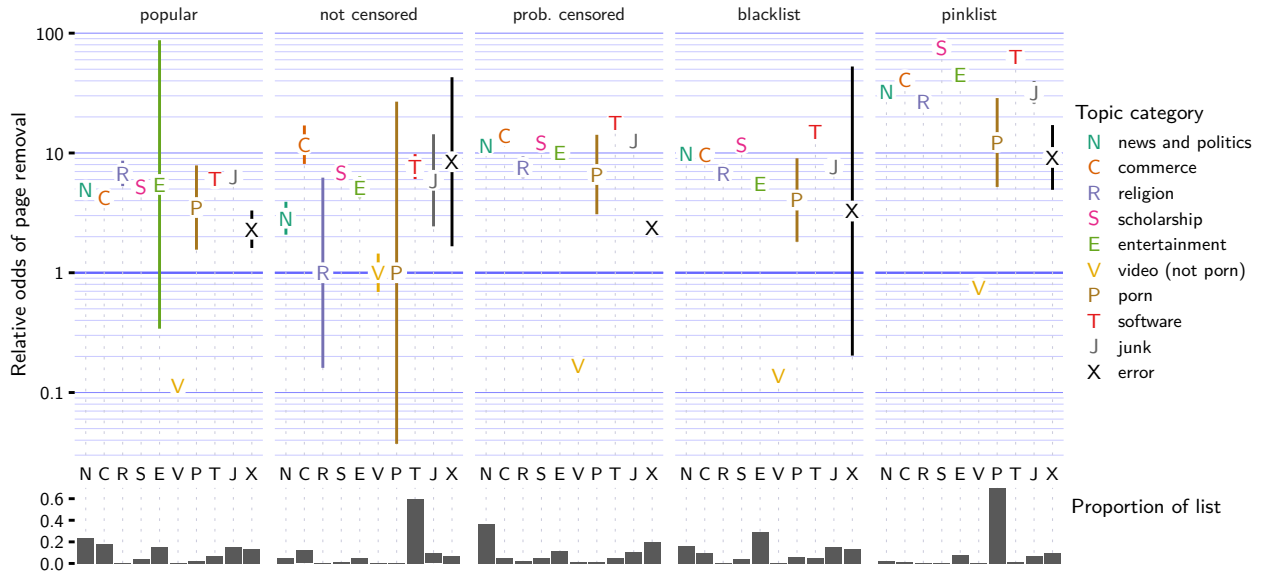

Fig. 5. How likely pages are to be taken down compared to Common Crawl pages. Error bars show 95% confidence intervals.

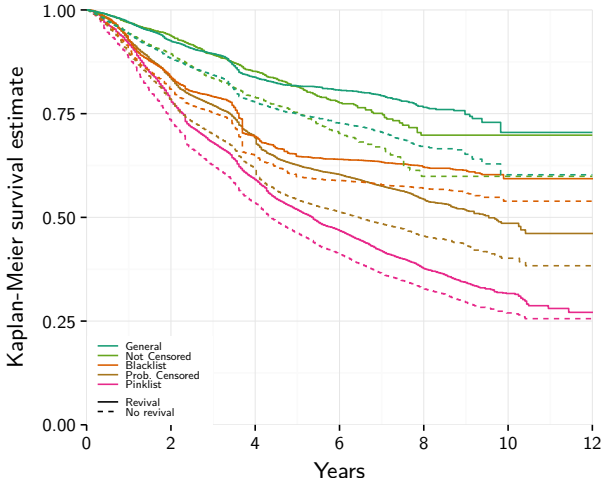

Fig. 6. Kaplan-Meier curves for different lists (best viewed in color).

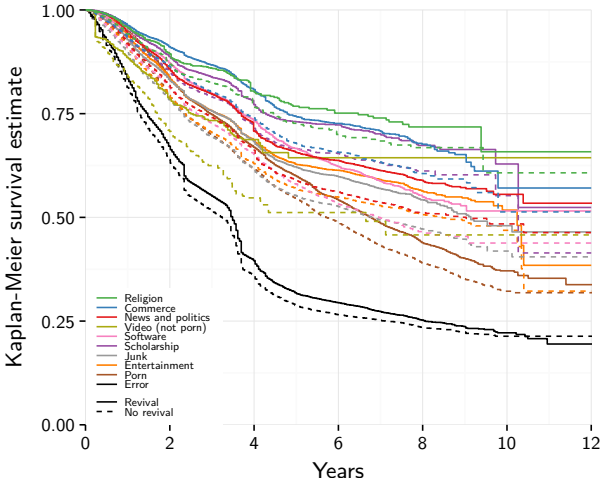

Fig. 7. Kaplan-Meier curves for different categories of topic (best viewed in color).

short lifetime of the "error" category reflects that once pages start turning into error messages, the entire site is likely to go away.) This in itself demonstrates the need for frequent updates to probe lists.

To reveal how lists and topics interact in determining page lifetime, we use a two-variable Cox proportional hazard model:

$$h_i(t) = h_0(t)e^{\beta_1 L_i + \beta_2 T_i} \tag{3}$$

where L_i is the type of page (blacklisted, etc), T_i is the topic category, h_0, β_1, and β_2 regression coefficients, and h_i the hazard rate at time t. Using this model, Figure 5 compares the odds of death of each type and category of page with those in the Common Crawl list. Each panel of this figure compares a group of source lists to Common Crawl; within each panel, there is a letter indicating the odds ratio, with 95% confidence interval, for each group of topics. Larger numbers on the y-axis indicate greater chances of a page being removed from the net. When the confidence interval is large, this means either that we have very little data (for instance, the "not censored" list group has fewer than 10 sites in the "religion" and "porn" categories) or that the lifetimes of pages in some category vary widely (for instance, "popular/entertainment").

This analysis confirms that, regardless of their topic, pages listed on the pinklists are more likely to be removed than pages on any other set of lists. Popular pages are somewhat less likely to be removed, but not as much as one would expect; this is probably because these lists include a fair number of sites that were only popular for the proverbial fifteen minutes. Curiously, non-pornographic video is *less* likely to be taken down than anything else; this may reflect the durability of video-sharing sites, which have to make a substantial infrastructure investment just to get started.

Another curious result is that, while pinklisted porn has a very short lifetime ($\beta_1 = 0.715$, $p < 0.01$), porn in general has a *longer* lifetime than the average ($\beta_2 = -0.123$, $p < 0.01$) (page revival allowed in both cases). This may reflect a fundamental dichotomy within the category. Legal pornography (in the USA) is a large and well-funded industry, capable of maintaining stable websites for years on end. However, there is also a substantial gray market, consisting of many small-time operations looking to make money fast and not so much concerned with law or ethics. These sites turn over very quickly indeed, and are perhaps also more likely to come to the negative attention of a censorship bureau.

8 Conclusions

We performed a large-scale study of 758,191 unique URLs drawn from 22 sources, developing for the purpose a system that can automatically retrieve and analyze this volume of text. We compared lists of pages reported to be censored with multiple control lists. We found that these lists are not easily compared directly, but patterns emerge when the pages are downloaded and analyzed for their topics. Cross-country patterns of censorship are readily detectable by comparing topics; pornography features prominently, but so do social media, music (copyright infringement?), and regional news. Survival analysis of web pages within each topic and each source provides convincing evidence that potentially controversial pages tend to have shorter lifetimes than less sensitive pages. The topic of a page is a significant predictor of its lifespan, and appearing on certain types of lists is also an effective predictor, even when controlling for topic.

8.1 Building Better Probe Lists

From our measurements, a number of guidelines emerge on how to build better probe lists for automated, at-scale measurement of Internet censorship. To achieve both depth and breadth of coverage, one should start with a topic balance across webpages of interest that is consistent with the web at large. If a censor is known to object to specific topics, these topics may deserve to be weighted more heavily; however, the odds of noticing censorship are proportional to the size of the *intersection* of the probe list with the blacklist. Thus, when a censor attempts to block a given topic comprehensively, probe lists need not weight that topic heavily.

Most of the blacklists are heavily weighted toward topics relevant to the countries they came from. This is a point in favor of ONI's split list design, with one set of URLs to test everywhere and then additional sets to test in each country. However, the rapid decay of controversial pages demonstrates that it is imperative to update probe lists frequently.

Our crawler and topic analyzer could serve as the basis for a system that continually generates and refines probe lists with minimal human effort. The topic model can be used to select keywords to use in searching the web for newly created pages on each topic, thus ensuring freshness and depth, while reducing manual effort. It could also, be applied to identify keywords that fill gaps in the topic space, thus improving breadth as well. Finally, by comparing the topic coverage of a list to a sample of the Web at large, situations where a list has too much coverage of some topic or topics can be identified, improving efficiency.

8.2 Modeling Refinements

This study demonstrates the power of natural language processing to reason about the contents of collected webpages, even using crude approximations such as bag-of-words document representations and dictionary-lookup translation. However, context-aware translation would almost certainly improve our classification, considering that people often use metaphors and ellipses to get around keyword blacklists [19]. We can also refine our topic model by using information which is already collected but not analyzed, such as words found in the URL of the site and in its outbound links. And the "web page boilerplate" and "error message" topics demonstrate that our various preprocessing heuristics could still be improved.

The large number of languages present in our data set poses unique challenges. 11.5% of all the words were either unrecognized by CLD2, untranslatable by Google, or both. This is a larger fraction of the data set than any single language other than English. Some of it is non-words (e.g. strings of symbols and numbers) but, obviously, more comprehensive language resources would be better. In addition, segmentation tools are not available for all of the languages that are written without spaces between the words. In our data set, the most prominent lacuna is Tibetan.

If the legal obstacles can be resolved, augmenting the topic model with information from images might be an interesting experiment. The state of the art in machine classification of images is well behind that for text, but is advancing rapidly. We suspect that many of the presently unclassifiable pages, especially those where no text survives boilerplate removal, are image-centric.

Finally, our statistical analysis of topic correlation with sources relies on the assumption that Common Crawl is topic-neutral. Unfortunately, Common Crawl is more strongly biased toward English than most of our other sources (see Appendix B) so this assumption is suspect. Alternatives exist, but they have their own deficiencies. For example, one can uniformly sample hostnames from the top-level DNS zones, but this only discovers website front pages. Developing a better "uniform" sample of the Web may well be a project in itself.

9 Acknowledgments

This research was partially supported by the National Science Foundation (under awards CCF-0424422 and CNS-1223762); and by the Department of Homeland Security Science and Technology Directorate, Cyber Security Division (DHS S&T/CSD), the Government of Australia and SPAWAR Systems Center Pacific (through BAA-11.02, contract number N66001-13-C-0131). This paper represents the position of the authors and not that of the aforementioned agencies.

We thank Luís Brandão, Pamela Griffith, Ariya Hidayat, Aya Kunimoto, Karen Lindenfelser, Srujana Peddada, Riana Pfefferkorn, and Kyle Soska for technical and editorial assistance in the preparation of this paper.

References

[1] Nicholas Aase, Jedidiah R. Crandall, Álvaro Díaz, Jeffrey Knockel, Jorge Ocaña Molinero, Jared Saia, Dan Wallach, and Tao Zhu. "Whiskey, Weed, and Wukan on the World Wide Web: On Measuring Censors' Resources and Motivations." *Free and Open Communications on the Internet*. USENIX. 2012.

[2] Sarker Tanzir Ahmed, Clint Sparkman, Hsin-Tsang Lee, and Dmitri Loguinov. "Around the Web in Six Weeks: Documenting a Large-Scale Crawl." *INFOCOM*. IEEE. 2015, pp. 1598–1606.

[3] Collin Anderson, Philipp Winter, and Roya. "Global Network Interference Detection over the RIPE Atlas Network." *Free and Open Communications on the Internet*. USENIX. 2014.

[4] Daniel Anderson. "Splinternet Behind the Great Firewall of China." *ACM Queue* 10.11 (2012).

[5] Luu Tuan Anh and Kazuhide Yamamoto. *DongDu — Vietnamese Word Segmenter*. Software library. 2012.

[6] Simurgh Aryan, Homa Aryan, and J. Alex Halderman. "Internet Censorship in Iran: A First Look." *Free and Open Communications on the Internet*. USENIX. 2013.

[7] Ziv Bar-Yossef and Sridhar Rajagopalan. "Template Detection via Data Mining and Its Applications." *World Wide Web*. ACM, 2002, pp. 580–591.

[8] Berkman Center for Internet and Society. *Herdict: help spot web blockages*. Web site.

[9] Chris Biemann, Felix Bildhauer, Stefan Evert, Dirk Goldhahn, Uwe Quasthoff, Roland Schäfer, Johannes Simon, Leonard Swiezinski, and Torsten Zesch. "Scalable Construction of High-Quality Web Corpora." *Journal for Language Technology and Computational Linguistics* 28.2 (2013), pp. 23–59.

[10] Steven Bird, Edward Loper, and Ewan Klein. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

[11] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (2003), pp. 993–1022.

[12] Jordan L. Boyd-Graber and David M. Blei. "Multilingual Topic Models for Unaligned Text." *Uncertainty in Artificial Intelligence*. Ed. by Jeff A. Bilmes and Andrew Y. Ng. AUAI Press, 2009, pp. 75–82.

[13] Leo Breiman. "Random Forests." *Machine Learning* 45.1 (2001), pp. 5–32.

[14] Sam Burnett and Nick Feamster. "Encore: Lightweight Measurement of Web Censorship with Cross-Origin Requests." *SIGCOMM*. ACM. 2015.

[15] Abdelberi Chaabane, Terence Chen, Mathieu Cunche, Emiliano De Cristofaro, Arik Friedman, and Mohamed Ali Kaafar. "Censorship in the Wild: Analyzing Internet Filtering in Syria." *Internet Measurement Conference*. ACM. 2014.

[16] Sharma Chakravarthy and Subramanian C Hari Hara. "Automating Change Detection and Notification of Web Pages." *Database and Expert Systems Applications*. IEEE. 2006, pp. 465–469.

[17] Pi-Chuan Chang, Michel Galley, and Christopher D Manning. "Optimizing Chinese Word Segmentation for Machine Translation Performance." *Statistical Machine Translation*. Association for Computational Linguistics. 2008, pp. 224–232.

[18] Richard Clayton, Steven J. Murdoch, and Robert N. M. Watson. "Ignoring the Great Firewall of China." *Privacy Enhancing Technologies*. 2006, pp. 20–35.

[19] Jedidiah R. Crandall, Daniel Zinn, Michael Byrd, Earl Barr, and Rich East. "ConceptDoppler: A Weather Tracker for Internet Censorship." *Computer and Communications Security*. ACM. 2007.

[20] Jakub Dalek, Bennett Haselton, Helmi Noman, Adam Senft, Masashi Crete-Nishihata, Phillipa Gill, and Ronald J. Deibert. "A Method for Identifying and Confirming the Use of URL Filtering Products for Censorship." *Internet Measurement Conference*. ACM. 2013, pp. 23–30.

[21] Ronald Deibert, John Palfrey, Rafal Rohozinski, and Jonathan Zittrain, eds. *Access Denied: The Practice and Policy of Global Internet Filtering*. ONI Access 1. MIT Press, 2008.

[22] Ronald Deibert, John Palfrey, Rafal Rohozinski, and Jonathan Zittrain, eds. *Access Controlled: The Shaping of Power, Rights, and Rule in Cyberspace*. ONI Access 2. MIT Press, 2010.

[23] Ronald Deibert, John Palfrey, Rafal Rohozinski, and Jonathan Zittrain, eds. *Access Contested: Security, Identity, and Resistance in Asian Cyberspace*. ONI Access 3. MIT Press, 2011.

[24] Marcel Dischinger, Massimiliano Marcon, Saikat Guha, Krishna P. Gummadi, Ratul Mahajan, and Stefan Saroiu. "Glasnost: Enabling End Users to Detect Traffic Differentiation." *Networked Systems Design and Implementation*. USENIX. 2010.

[25] Stefan Evert. "A lightweight and efficient tool for cleaning Web pages." *International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA), 2008.

[26] Arturo Filastò and Jacob Appelbaum. "OONI: Open Observatory of Network Interference." *Free and Open Communications on the Internet*. USENIX. 2012.

[27] Sean Gallagher. "Big Brother on a budget: How Internet surveillance got so cheap." *Ars Technica* (2012).

[28] Phillipa Gill, Masashi Crete-Nishihata, Jakub Dalek, Sharon Goldberg, Adam Senft, and Greg Wiseman. "Characterizing Web Censorship Worldwide: Another Look at the OpenNet Initiative Data." *ACM Transactions on the Web* 9.1 (2015).

[29] Google Inc. *Translate API*. Web service.

[30] Joseph C. Gratz, Marvin Ammori, and Lavon Ammori. "Brief of *amici curiae* Automattic Inc.; Google Inc.; Twitter Inc.; and Tumblr, Inc." *Lenz v. Universal Music*. 801 F.3d 1126. (9th Cir. 2015).

[31] Joseph Henrich, Steven J. Heine, and Ara Norenzayan. "The weirdest people in the world?" *Behavioral and Brain Sciences* 33.2–3 (June 2010), pp. 61–83.

[32] Ariya Hidayat et al. *PhantomJS*. Software application.

[33] Internet Archive. *Wayback Machine*. Web service.

[34] Ben Jones, Tzu-Wen Lee, Nick Feamster, and Phillipa Gill. "Automated Detection and Fingerprinting of Censorship Block Pages." *Internet Measurement Conference*. ACM. 2014.

[35] Edward L. Kaplan and Paul Meier. "Nonparametric Estimation from Incomplete Observations." *Journal of the American Statistical Association* 53 (1958), pp. 457–481.

[36] Theppitak Karoonboonyanan, Pattara Kiatisevi, Vuthichai Ampornaramveth, Poonlap Veerathanabutr, and Chanop Silpa-Anan. *LibThai*. Software library. 2001–2013.

[37] Sheharbano Khattak, Mobin Javed, Syed Ali Khayam, Zartash Afzal Uzmi, and Vern Paxson. "A Look at the Consequences of Internet Censorship Through an ISP Lens." *Internet Measurement Conference*. ACM. 2014.

[38] Jeffrey Knockel, Jedidiah R Crandall, and Jared Saia. "Three Researchers, Five Conjectures: An Empirical Analysis of TOM-Skype Censorship and Surveillance." *Free and Open Communications on the Internet*. USENIX. 2011.

[39] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. "Applying Conditional Random Fields to Japanese Morphological Analysis." *Empirical Methods in Natural Language Processing*. ACL, 2004, pp. 230–237.

[40] Shuang Lin, Jie Chen, and Zhendong Niu. "Combining a Segmentation-Like Approach and a Density-Based Approach in Content Extraction." *Tsinghua Science and Technology* 17.3 (2012), pp. 256–264.

[41] Andrew K McCallum. *{MALLET: A Machine Learning for Language Toolkit}*. Software library. 2002.

[42] David M. Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. "Polylingual Topic Models." *Empirical Methods in Natural Language Processing*. ACL, 2009, pp. 880–889.

[43] Will Monroe, Spence Green, and Christopher D. Manning. "Word Segmentation of Informal Arabic with Domain Adaptation." *ACL Short Papers* (2014).

[44] Zubair Nabi. "The Anatomy of Web Censorship in Pakistan." *Free and Open Communications on the Internet*. USENIX. 2013.

[45] Zubair Nabi. "Censorship is Futile." *First Monday* 19.11 (2014).

[46] David Newman, Arthur U. Asuncion, Padhraic Smyth, and Max Welling. "Distributed Algorithms for Topic Models." *Journal of Machine Learning Research* 10 (2009), pp. 1801–1828.

[47] Jong Chun Park and Jedidiah R. Crandall. "Empirical Study of a National-Scale Distributed Intrusion Detection System: Backbone-Level Filtering of HTML Responses in China." *Distributed Computing Systems*. IEEE. 2010, pp. 315–326.

[48] M. F. Porter. "An Algorithm for Suffix Stripping." *Readings in Information Retrieval*. Ed. by Karen Sparck Jones and Peter Willett. San Francisco: Morgan Kaufmann, 1997, pp. 313–316.

[49] *Readability*. Web service.

[50] Caroline R. Richardson, Paul J. Resnick, Derek L. Hansen, Holly A. Derry, and Victoria J. Rideout. "Does Pornography-Blocking Software Block Access to Health Information on the Internet?" *The Journal of the American Medical Association* 288.22 (2002), pp. 2887–2894.

[51] Stephen Robertson. "Understanding inverse document frequency: on theoretical arguments for IDF." *Journal of Documentation* 60.5 (2004), pp. 503–520.

[52] Andreas Sfakianakis, Elias Athanasopoulos, and Sotiris Ioannidis. "CensMon: A Web Censorship Monitor." *Free and Open Communications on the Internet*. USENIX. 2011.

[53] Dick Sites. *Compact Language Detection 2*. Software library. 2013–.

[54] Kyle Soska and Nicolas Christin. "Automatically Detecting Vulnerable Websites Before They Turn Malicious." *USENIX Security Symposium*. 2014, pp. 625–640.

[55] Ramesh Subramanian. "The Growth of Global Internet Censorship and Circumvention: A Survey." *Communications of the IIMA* 11.2 (2011).

[56] Fei Sun, Dandan Song, and Lejian Liao. "DOM Based Content Extraction via Text Density." *Research and Development in Information Retrieval*. ACM, 2011, pp. 245–254.

[57] Janos Szurdi, Balazs Kocso, Gabor Cseh, Jonathan Spring, Mark Felegyhazi, and Chris Kanich. "The Long "Taile" of Typosquatting Domain Names." *USENIX Security Symposium*. 2014, pp. 191–206.

[58] The Common Crawl Foundation. *Common Crawl*. Web site.

[59] The OpenNet Initiative. *ONI Country Profiles*. Web site.

[60] The OpenNet Initiative. *United States and Canada Overview [of Censorship]*. Web report.

[61] The OpenNet Initiative. *URL testing lists*. Git repository.

[62] The OpenNet Initiative. *Country profiles: Thailand*. Web report. 2012.

[63] John-Paul Verkamp and Minaxi Gupta. "Inferring Mechanics of Web Censorship Around the World." *Free and Open Communications on the Internet*. USENIX. 2012.

[64] Thomas Vissers, Wouter Joosen, and Nick Nikiforakis. "Parking Sensors: Analyzing and Detecting Parked Domains." *Network and Distributed System Security Symposium*. Internet Society. 2015.

[65] Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David M. Mimno. "Evaluation methods for topic models." *International Conference on Machine Learning*. Ed. by Andrea Pohoreckyj Danyluk, Léon Bottou, and Michael L. Littman. Vol. 382. ACM, 2009, pp. 1105–1112.

[66] David Y. Wang, Stefan Savage, and Geoffrey M. Voelker. "Cloak and Dagger: Dynamics of Web Search Cloaking." *Computer and Communications Security*. ACM. 2011, pp. 477–490.

[67] Joss Wright. "Regional Variation in Chinese Internet Filtering." *Information, Communication & Society* 17.1 (2014), pp. 121–141.

[68] Joss Wright, Tulio de Souza, and Ian Brown. "Fine-Grained Censorship Mapping: Information Sources, Legality and Ethics." *Free and Open Communications on the Internet*. USENIX. 2011.

[69] Pengtao Xie, Yuntian Deng, and Eric Xing. "Diversifying Restricted Boltzmann Machine for Document Modeling." *Knowledge Discovery and Data Mining*. ACM. 2015, pp. 1315–1324.

[70] Xueyang Xu, Z. Morley Mao, and J. Alex Halderman. "Internet Censorship in China: Where Does the Filtering Occur?" *Passive and Active Measurement*. 2011, pp. 133–142.

[71] Limin Yao, David M. Mimno, and Andrew McCallum. "Efficient methods for topic model inference on streaming document collections." *Knowledge Discovery and Data Mining*. Ed. by John F. Elder IV, Françoise Fogelman-Soulié, Peter A. Flach, and Mohammed Javeed Zaki. ACM, 2009, pp. 937–946.

[72] Taha Yasseri, Anselm Spoerri, Mark Graham, and János Kertész. "The Most Controversial Topics in Wikipedia: A multilingual and geographical analysis." *Global Wikipedia: International and cross-cultural issues in online collaboration*. Ed. by Pnina Fichman and Noriko Hara. Rowman & Littlefield, 2014.

[73] Tao Zhu, David Phipps, Adam Pridgen, Jedidiah R. Crandall, and Dan S. Wallach. "The Velocity of Censorship: High-Fidelity Detection of Microblog Post Deletions." *USENIX Security Symposium*. USENIX. 2013, pp. 227–240.

A Evaluation of Parked Domain Detection Algorithms

Vissers et al. [64] developed a random-forest classifier [13] based on features extracted from the HTML and the HTTP transactions at page load time. It relies on structural differences between a parked domain and a normal domain, such as the ratio of text to markup, the ratio of internal to external links, and the number of nested pages ("frames"). Szurdi et al. [57], developed a set of regular expressions based on the templates used by specific domain parkers, while investigating the related practice of *typosquatting*. Typosquatters place often-malicious sites at misspellings of the names of popular websites, e.g. googel.com for Google.

We evaluated our classifiers on three data sets, "PS," "LT," and "Cen." PS is the set used by Vissers et al. [64] to assess their classifier. It includes 3,047 non-parked domains taken from Alexa (see Section 3), and 3,227 parked domains operated by 15 parkers. LT was used by Szurdi et al. [57] to evaluate their classifier. It consists of 2,674 pages collected from typosquatted domains, and manually labeled; 996 are parked and 1,678 not parked. Finally, Cen consists of 100 pages randomly selected from our contemporary data.

To train the random-forest classifier, we combined PS and LT, and then split the combination 80/20 for training and testing. Neither LT nor Cen includes HTTP transaction information, so the features depending on this data were disabled. Despite this, we reproduce Vissers et al.'s results on PS, which indicates that those features are not essential. The regexp classifier does not

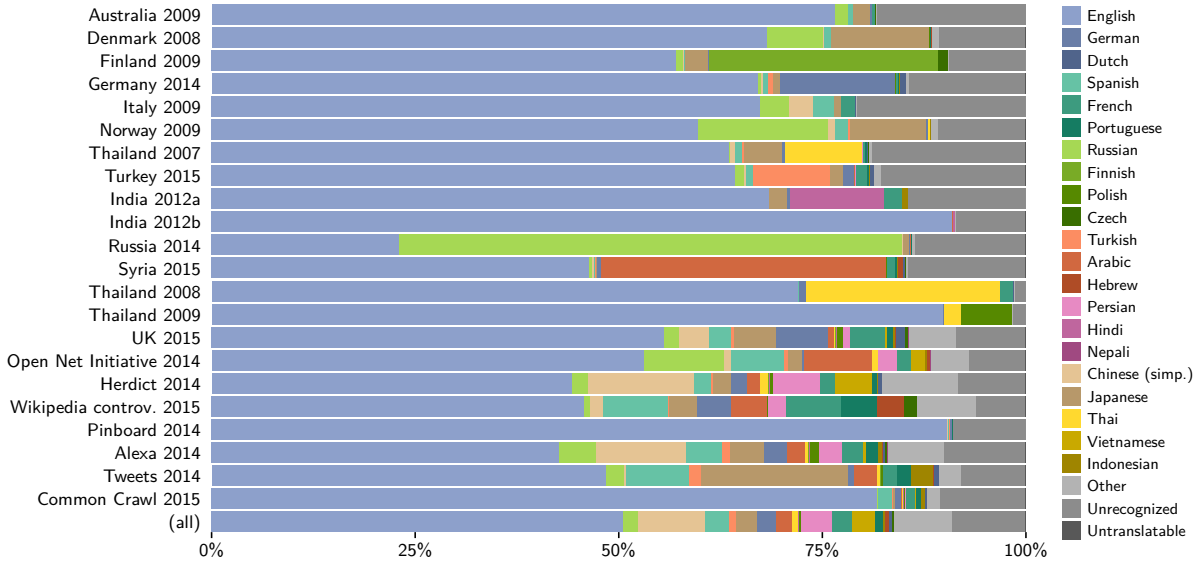

Fig. 8. The proportion of each source list devoted to text in each of the most common 21 languages.

require training, but we augmented the original battery of regular expressions with new rules derived from PS.

Table 4. Performance of the two parked-domain detectors on three data sets.

Algorithm	Random forest			Regexps		
Dataset	PS	LT	Cen	PS	LT	Cen
Accuracy	97.9	93.1	89.0	95.0	89.6	99.0
Precision	99.2	89.9	42.9	99.9	96.9	100.0
Recall	96.9	92.1	30.0	90.4	79.1	90.0

Table 4 shows the performance of both classifiers on all three datasets. The random-forest classifier performs reasonably well on PS and LT, with precision and recall both 90% or above, but poorly on Cen: precision drops to 42.9% and recall to 30.0%. (Accuracy remains high because Cen is skewed toward non-parked pages.) The (improved) regular-expression classifier, on the other hand, performs well on all three; its worst score is 79.1% recall for LT.

To better understand why the random-forest classifier performs poorly on Cen, we constructed a larger version of it containing 7,422 pages. Both classifiers agreed on 6,869 of these: 81 parked, 6,788 not parked. 447 pages were classified as parked only by the regular-expression classifier, and 106 pages only by the random-forest classifier. We manually verified a subsample of 25 pages in each category. In all cases, the regular-expression classifier was correct; where they disagreed, the random-forest classifier was invariably wrong. The most common cause of errors was pages using frames to load most of their content. The random-forest classifier treats this as a strong signal that the page is parked, but this is inaccurate for Cen.

B Language Biases of Sources

Figure 8 shows, for each source list, what proportion of its non-boilerplate text is in each of the most commonly used 21 languages (plus "other," "unrecognized," and "untranslatable"). English, unsurprisingly, dominates nearly all of the lists—the surprise is when it doesn't, as in the Russian blacklist. We suspect this might also occur for Chinese, if we had a Chinese blacklist. Where a single language dominates non-English text, it is also unsurprising: German for Germany, Russian for Russia, Arabic for Syria, Thai for Thailand. ONI, Herdict, Wikipedia, Alexa and Twitter show no dominant second language, again as expected.

Four lists have hardly anything *but* English. India 2012b and Thailand 2009 are dominated by videos posted to YouTube and similar sites, for which the common language for leaving comments is English; the videos themselves may well have featured other languages. Most of these videos have since been taken down, so we could not spot-check them. Pinboard, the negative control, was compiled by someone who is only fluent in English. This may mean that our topic model is largely

ignorant of *innocuous* material in languages like Russian and Arabic, but this is harmless for now. It will become a problem in the future, when we attempt to make automated judgements about whether pages are worth including in a continuously updated probe list.

Finally, the dominance of English in the Common Crawl data is inexplicable and disturbing. It may indicate a methodological error, either on the part of Common Crawl itself, or in our selection of a subsample. One possibility is that there are too few *sites* in our subsample, and those sites are largely Anglophone. Another is that their crawler may not have started from the right "seed" locations within the hyperlink graph to find much material in other languages. Regardlesss of the cause, though, this casts doubt on our assumption that this subsample is a good baseline for cross-list comparison. Anything that is in other languages may seem more unusual than it is, by comparison.