This chapter is dedicated to rationales and research techniques to investigate state internet censorship at a juncture when the rationales remain clear, but the techniques may be ethically objectionable, owing to surveillance regimes in place in the countries under study. How to carry on with the important work on censorship detection and monitoring without putting the researchers and/or the computer users in the countries in question in harm’s way? In the following, research questions for censorship inquiries are laid out, such as its discovery, unintended consequences, effectiveness, circumvention, censor duping, Western complicity, censored content effects as well as the effects of being watched (self-censorship). Additionally, techniques are put forward for drawing up URL (and keyword) lists to be studied, as well as for fetching URLs in the countries in question to check for blockage. Short (historical) case studies concerning Chinese and Iranian internet censorship illustrate certain lines of inquiry and as well as techniques, including current ethical issues. The chapter concludes with considerations about practising and contributing to internet censorship research when its ethical and technical study are at a crossroads.

The evolution of state internet censorship and its study

Generally speaking, internet censorship research rests on sourcing claims about existing blocking practices for the country or countries under study, building URL or website lists, and checking the websites in the countries for censorship. When a website is censored by the state (such as in Iran), a so-called block page appears, with an official message alerting the user to the illegality of accessing the content (see Figure 4.1). With the block page or a
less overt technical indication in countries that do not use them (such as in China where the connection times out), the website is confirmed as censored. Censorship practices online subsequently may be indexed, reported and interpreted.

Over the past few years, however, internet censorship has become more involved, and with it its study. Not just websites as such are blocked, in the way a book or film in the past was banned (Darnton, 2014). Keywords may be targeted, such as those appearing in URLs, search engine queries or content typed into messaging services. One case in point occurs in Weibo and WeChat in China, where a long list of words relating to a particular government crackdown returns no results in the one, and messages containing the words remain undelivered in the other (Ruan et al., 2017).

There is a certain technicity to how content is censored on the internet. Websites may be blocked temporarily or only in certain cities like the capital during elections or a street protest; such spatial and temporal blocking is distinctive from national content bans for all citizens and has been dubbed second-generation internet censorship. Internet connection speeds may be slowed, or ground to a halt, as took place in Iran in June 2009 during the election crisis. Websites loading in the browser may time out or be pointed to the wrong address. When users are not only blocked from particular websites (through covert means, without a block page appearing) but also their requests are monitored, censorship may be coupled with surveillance and user logging, or what researchers have called third-generation internet censorship (Deibert and Rohozinski, 2010).

How to capture and study the phenomenon? Are there also techniques to capture these first, second as well as third generations of internet censorship? Are there other forms that do not fit the categories? For countries that block URLs (and perhaps put up censorship notices), simple techniques of URL fetching in a browser on location, or remotely through proxies or virtual private networks (VPNs), allow the researcher to gain a sense of the kind of content being censored. Repeated requests would show the persistence of censorship, and likely content policies. But for the so-called next-generation types of censorship (which do not necessarily follow chronologically from one another and may occur simultaneously), updated censorship detection tools and research practices are required. Temporary and/or regional blocking demands geographically and temporally distributed monitoring and takes more time and resources. It becomes necessary to use multiple researchers, volunteers, proxies and VPNs with known locations using IP-to-geo look-up services to pinpoint the server coordinates. It also may become important to run more voluminous URL requests through the servers in the countries in question, given that blocking may be done ‘just in time’. One could imagine that as the volume of queries increases the prospect of those being flagged by system administrators as unusual behaviour (possibly to be reported) may increase.
Figure 4.1  Block page in Iran redirects to a directory of acceptable national websites.

The ethical stakes

Discussing internet censorship research that relies on third parties (without their consent or without proper discussion of potential repercussions), be they friendly volunteers who have agreed to run some URLs, or the owners of proxy servers, one researcher writes that ‘a user innocently aiding a researcher in mapping their national filter, resulting in their computer suddenly attempting to connect to all forms of banned content, may find themselves under very unwelcome scrutiny’ (Wright et al., 2011: 5). In a later article, a similar point is made less equivocally: ‘Deliberate misuse of a network service for the purposes of detecting internet filtering may be illegal in many jurisdictions, and such misuse without the consent of users or system operators is clearly unethical’ (Wright, 2014: 128). Raising the ethical stakes for researchers performing remote internet censorship research (and repurposing devices in order to do so) also coincide with the awareness of the mass state surveillance reported by Edward Snowden (beginning in June 2013), and the spreading of third-generation censorship practices to more countries through Western companies such as the Hacking Team, based in Milan. Concerns about the state watching internet censorship research raise questions. How to monitor and record internet censorship when the means to do so are (potentially) illegal in the country in question? Would such research that uses a third party – without their knowing it – put that person at risk? How to mitigate, or eliminate, the risks?

Figure 4.2 Global internet filtering map by the Open Net Initiative, last updated December 2014, indicating that countries have pervasive, substantial, selective, suspected or no evidence of blocking websites.

Source: http://map.opennet.net/filtering-pol.html.
As stated at the outset, it is of importance to be informed of the censorship practices in the country under question. Are they known to practise next-generation monitoring and surveillance? If so, what kinds of censorship research practices would still be considered responsible? Joss Wright developed a technique to test URLs for filtering that did not involve third parties in China; it relies on a particular censorship practice (DNS hijacking/poisoning/tampering) and thus is well suited to China but less so to the countries using other techniques such as filtering software purchased from Western companies. Other researchers engage informed volunteers and/or employ stealth DIY research practices, running URLs through servers inside the countries in question and securely exporting the data. (This research protocol was performed for nearly a decade in the pioneering work at the Open Net Initiative.) Papers are published with pseudonyms, detailing how the investigative work was done in-country, without tipping off the authorities (Aryan et al., 2013). Still others use proxies and VPNs, with or perhaps without some grasp of the origins of these services and the people behind them.

Here the focus is on the kinds of rationales put forward for testing for blocking, generally, given the sensitivity of (repeatedly) requesting banned websites in some countries. Additionally, there is the question of which internet censorship detection technique may be used for different types of content under scrutiny. As a case in point, majority Muslim countries are known to block alcohol-related websites, so sporadic and minimal testing could be performed to confirm that finding (if deemed worthy of research), rather than systematic and voluminous checks (Noman, 2011). Thus, here I concentrate on ways in which to avoid having research experience a chilling effect by its object of study. The emphasis is on the capacity for observing and monitoring internet filtering and the manner in which to do so responsibly.

**Doing internet censorship research with research questions**

In describing techniques for building URL lists, the aim is also to formulate compelling research questions about internet censorship, beginning with the ethical ones (as above) but also with respect to its discovery, unintended consequences, effectiveness, circumvention, censor duping and effects on the content being censored or its authors being watched (self-censorship). There is also detective work in tracing the filtering software used by censoring countries to its Western manufacturer. Did a US (or Canadian, British, German, French or Italian) company really sell censorship software to those authoritarian countries? How much credence can be lent to the claim that the companies are following the laws of the country to which they sell the software (Reporters without Borders, 2017)? Also described are tools for the use of proxies and VPNs, together with guidelines concerning how to decide when and how intensively to use them.
At the outset it may be instructive to mention early claims (in the 1990s) about the internet being immune to censorship, and contrast those with the general observations made two decades later by the most significant internet censorship research, the Open Net Initiative (Elmer-Dewitt, 1993; Rogers, 2009c). The ‘civil libertarian’s dream’ built into the architecture of the internet that would treat censorship as a ‘multifunction and route around it’ has been supplanted by ‘extensive [national] filtering practices’ (Boyle, 1997; Open Net Initiative, 2014). By the time it suspended its research at the end of 2014, the Open Net Initiative’s decade-long programme sought evidence of online censorship in some 60 countries (see Figure 4.2).

How to carry on with the work? This chapter provides some (modest) means to undertake internet censorship research by investigating URLs blocked by the state, where the examples are drawn from the more studied countries (such as China and Iran) that block or interrupt websites for their sensitive content (see Figure 4.3). URLs containing particularly troubling keywords may be filtered, as are some search results. Connections are also slowed or timed out.

Seemingly straightforward – in that content is or is not blocked – internet censorship research is actually rather technical, and the terminology used suffused with computational lingo, especially in the web science literature. For example, Iran is known to employ the following censorship techniques: ‘HTTP host-based blocking, keyword filtering, DNS hijacking, and protocol-based throttling’ (Aryan et al., 2013: 1). The first term refers to website blocking, and the second to the same (when a particular keyword appears in the

---

**Figure 4.3** Miao Ying and Mehdi Yahyanejad’s slide concerning leading websites blocked in China and Iran at Rhizome’s Seven on Seven conference, New Museum, New York City, 22 April 2017.

URL) (Warzel, 2013). Protocol-based throttling means slowing the connection speed, for example, to ‘https’ connections (secure protocol) because certain internet circumvention software employs it, among other reasons. ‘DNS hijacking’ describes how the web user (e.g., typing in “Facebook.com”) is sent to the wrong web location (i.e., to a numerical IP address other than Facebook’s).

Countries make large projects out of internet censorship, and NGOs and researchers monitor them, issuing reports and white papers that make (researchable) claims and provide justification for more work to be done to study censorship’s technicity and effects.

China and Iran

China is known to have a proverbial ‘Great Firewall’, known as the ‘Golden Shield Project’ (Dong, 2012). Chinese state filtering, however, is not often overt; neither a block page appears, nor a directory with a list of acceptable websites, as in the case of Iran (see Figure 4.1). As mentioned, banned websites (such as Facebook since 2009) time out, meaning that after attempting to load there is an unexpected disconnection. That tampering (rather than outright blocking) has a series of implications for internet users and researchers, the most crucial one of which is thought to be the possible logging of those requesting unacceptable websites. As one researcher has put it: ‘there is evidence that DNS requests for blocked services may be logged in the case of China’ (Wright, 2014: 128). ‘Blocked in China’, the website that performs live second-generation censorship research (returning filtering results from Beijing, Shenzhen, Inner Mongolia, Heilongjiang Province and Yunnan Province) treats the concern on an FAQ page about how to circumvent censorship:

VPNs are not strictly illegal in China. There’s no precedent we know of in which someone has been arrested for using one. The goal of authorities is to censor, not punish. That being said, using a VPN to break Chinese laws or publish anti-Communist rants could still get you in trouble. (blockedinchina.net, 2017)

The extent to which VPN access is maintained in future is in question, however (Haas, 2017). In Iran there was a time when it was widely reported that the state was building its own ‘halal internet’ (a term used by an Iranian official, Ali Aghamohammadi), which would be equivalent to a white list of websites that users could access, with the rest of the World Wide Web (or ‘haram internet’) blocked (Farivar, 2011; Rhoads and Fassihi, 2011). The makings of such an internet have been visible since at least the Iran election crisis of 2009 when the block page would redirect to a directory listing of approved national sites. After the Stuxnet virus disrupted nuclear progress in 2011, the official rationale for the national internet project was put this way: ‘Isolation of the clean Internet from the unclean portion will make it impossible to use the Internet for unethical and dirty businesses’ (Article 19, 2016). In the event, the phased roll-out of the national internet has been supplemented by
the ‘intelligent filtering’ project which updates techniques of state censorship. While more research is called for, it is being monitored by NGOs and others, which have detailed the coincidental blocking of websites including ones that reported city government corruption in Tehran (Reporters without Borders, 2016b; Small Media, 2017). As is its mandate, Article 19, the advocacy group, has been vocal about (recent) Iranian filtering, providing also a compelling reason overall to study internet censorship:

If [the national internet is] fully implemented, there is no doubt that it would fly in the face of international standards on freedom of expression. In particular, Article 19 of the ICCPR [signed by Iran] provides that the right to freedom of expression includes the individual’s freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers and through any other media of his choice. (Article 19, 2016)

State internet censorship is often couched as a rights issue, prompting but also justifying its study, even if illegal. Other rationales for its study include unintended consequences such as over-blocking or the so-called ‘Streisand effect’ which stipulates that blocking content brings more attention to it, and prompts its further spread (Masnick, 2015). From the perspective of awareness-raising (among the NGO activities discussed here), it would be of normative value to point to blocked sites so as to highlight their contents, as Amnesty International’s Irrepressible Info anniversary project (mentioned below) once did. There is also a circulationist thesis to be explored; does blocked content become leaked and spread like samizdat, or through other means such as content aggregation? Here the question concerns the very effectiveness or even raison d’être of blocking.

Research questions also may treat the effects of censorship on content and voice. Does censorship kill content? In a study conducted on Iran it was found that despite being censored, Iranian bloggers continued to write postings and keep their websites fresh, providing circumstantial evidence to the idea that there is active censorship circumvention, as well as readerships outside of Iran in the diaspora and elsewhere (Rogers et al., 2012; Alimardani, 2014).

Table 4.1 Categories in the pioneering ‘Global URL List’ for state internet censorship research by the Open Net Initiative, 2006.

<table>
<thead>
<tr>
<th>Alcohol</th>
<th>Major events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anonymizers</td>
<td>Medical</td>
</tr>
<tr>
<td>Blogging domains</td>
<td>Miscellaneous</td>
</tr>
<tr>
<td>Drugs</td>
<td>News Outlets</td>
</tr>
<tr>
<td>Dating</td>
<td>P2P</td>
</tr>
<tr>
<td>Email</td>
<td>Porn</td>
</tr>
<tr>
<td>Encryption</td>
<td>Provocative attire</td>
</tr>
</tbody>
</table>
In other countries such as Saudi Arabia and Bahrain, filtering software is installed by the state, and populated with URL blacklists, whitelists and/or keywords (Reporters without Borders, 2014). Internet censorship research has been directed at detecting which filtering software is used so as to out (and shame) the purveyor, which may be based in the USA or Canada. As Open Net Initiative researchers found, for example, SmartFilter and WebSense (both USA-based) and NetSweeper (Canada) have been deployed in countries that censor the Internet (Babcock and Freivogel, 2015). (SmartFilter, now owned by McAfee, denied selling it to Iran.) European firms later stepped into the market, too, prompting Reporters without Borders (2013) to call this the ‘era of digital mercenaries’. Here the research is of interest to investigative journalism and corporate social responsibility (among other discourses) where the question concerns the ethics and business of selling software to authoritarian regimes to censor critical content. The Milan-based firm, Hacking Team, is perhaps the most egregious (and recent) case of unethical selling, where their software is used by a variety of authoritarian regimes to track down dissidents. Internal email exchanges between company employees, themselves available through a hack, reveal the story of a company concerned with Italian export regulations and the recipient countries’ own laws, rather than the ethics of putting people in harm’s way (Currier and Marquis-Boire, 2015).

**Internet censorship research techniques: From URL list building to checking URLs**

Once one has looked into the claims concerning internet censorship in certain countries, and considered the research questions to pose, the work commences to build URL lists and check them. To start, five general approaches to building URL lists are sketched: editorial, crowdsourcing, search engine work, device studies and dynamic URL sampling. URL
list-building approaches also may be combined (and lists merged or concatenated). The approaches have been used by censorship researchers as well as NGOs seeking to raise awareness of the issue, as is detailed. Subsequently, means by which to check URLs in the countries under study are treated.

Editorial list building

Conventionally website directories have been used to build URL lists, such as those of Yahoo!, Dmoz.org or national ones such as startpagina.nl for the Netherlands. For more than a decade now, however, the search engine has replaced the human-edited directory as the main content organization device of the web. Manually edited expert lists for websites, with global categories, are gradually becoming extinct, as are ‘link lists’ on websites. Yahoo!’s seminal directory became defunct in December 2014, and Dmoz’s lasted until March 2017. Amnesty International no longer has a link list of human rights organizations. Choike, the Global South NGO portal, does not appear to have updated its directory of NGOs per category for some years now. The United Nations maintains lists of accredited NGOs per category, however much certain of the lists appear as PDFs without links to the websites, prompting manual work (extracting NGO names from the PDF, looking up URLs and perhaps making decisions concerning an organization’s main website given subsites with different languages).

For research into a single country, or cross-country comparison research, often a steady, so-called global list of URLs is built. This list contains URLs organized into categories that would be run for all countries under investigation, so as to be able to make comparisons in the form of types of sites blocked. This is the original method employed by the Open Net Initiative. The research began in the mid-2000s with an internet censorship monitoring programme that relied on a global list of URLs culled from Web directories such as Yahoo!’s and dmoz.org’s, with some 37 categories of website types, as well as on country-specific lists, put together by country and subject matter experts (see Table 4.1).

Where may one start to build an editorial list these days? With respect to borrowing URL lists from institutions that expertly categorize websites, the Webby Awards may be considered. In its promotional literature to express the weightiness of its award, the International Academy of Digital Arts and Sciences states that in 2017 there were 1,029,615,402 websites and 387 Webby Award winners, thus giving an indication of the number of categories they consider (which has increased greatly since its founding in 1996). One may also take up specific categories (such as ‘activism’), and plumb earlier winners, making a longer list of significant websites for each category.

Since the classic directories and link-list makers have discontinued their web librarian work, Wikipedia is increasingly becoming the rare, manual alternative for link lists. It is an editorial source providing lists and list-building opportunities, especially with its directory-like articles such as ‘List of human rights organizations’ (https://en.wikipedia.org/wiki/
List of human rights organizations). Language versions of Wikipedia also should be considered; the list of human rights organizations in the German Wikipedia is distinctive from that in the English-language one. Wikipedia even has a list of ‘websites blocked in China’ (with, as discussed below, Greatfire.org as its source) as well as a series of other countries. At the other end of the spectrum from the Webby awards, Wikipedia’s editorial style tends to be exhaustive as opposed to selective, leading to lengthy lists that the researcher may wish to cull by creating a significance threshold.

Wikipedia is also a source for keywords through the names of its articles or by other means. In one internet censorship research project, which one could call a baseline censorship ‘discovery’ project, the researchers collected all article titles from the Chinese language Wikipedia and searched for them on Weibo. Those queries returning the Weibo-service message saying ‘based on Chinese laws and regulations, search results are not shown’ were retained. Eventually the project compiled ‘Blocked on Weibo’, a list of words that are on Chinese censors’ blacklist. Here Wikipedia is used as a source, but querying the terms in Weibo reduces the otherwise long list.

Crowdsourcing blocked sites

A second technique for building URL lists is crowdsourcing. Through this approach, a list of websites is sourced from the outside by asking others, for example, to fill in online forms or install an extension. To that end, the expertise is also outsourced. It is a technique that has particular rationales for being undertaken, such as the sheer scale of the work to be done or the challenge of foreign languages. Seminal cases of crowdsourcing, borrowed from earlier resource sharing projects such as SETI at home, have included reCaptcha and Google Image Labeller, where the former has users help to transform scanned text into machine-readable material and the latter with tagging images or checking whether machine techniques are working.

Crowdsourcing has been applied to internet censorship awareness-raising as well as research. One of the earliest cases of awareness-raising is greatfirewallofchina.org, which asked the user to input his or her own website (or whichever one the user chooses), whereupon it would test it for blocking in China, thereby gradually building up a database with a history of discovered or confirmed blocks. In its earliest version, if the website were blocked, it became a kind of brick in a visual display of a great wall of China. One also could download and display a banner on one’s website, indicating that it has been tested. Another project with a similar name (greatfire.org) is less artistic and awareness oriented, and more service-minded with research capacities. Greatfire.org provides means to circumvent censorship, while also monitoring multiple online platforms for the state of Chinese censorship more generally (see Figure 4.4). One also may test websites as well as keywords. While one may test sites for blocking, greatfire.org has what I would call a ‘device studies’ approach (using multiple online platforms) rather than primarily a crowdsourced one (multiple users), as is discussed below.
Another early example of crowdsourcing was irrepressible.info, built for Amnesty International’s 45th anniversary in 2006. It showed content from websites that were blocked, making it a blocked content collection website, with show-and-tell capacities, so to speak. It thereby also took a circulationist approach to censorship circumvention in so far as one could place content from blocked websites on irrepressible.info, and thereby make it available (and thus ‘irrepressible’), though of course irrepressible.info itself could be blocked. Herdict.org (the verdict of the herd in contrast to the wisdom of the crowd), the project by the Berkman Center at Harvard University, also asks the crowd for URLs (Palfrey and Zittrain, 2011). One becomes a user of Herdict by downloading the Firefox add-on and providing one’s location. Then, when blocked websites are encountered, they can be reported to Herdict, together with one’s location. Individual URLs also can be entered in order to find out the status of those particular URLs per user location. In this way, the herd is meant to visit more websites (than any individual could) and ultimately compile a much larger reporting on the state of internet censorship. Perhaps more to the point, one may also use Herdict’s lists of blocked websites so as to confirm or otherwise chart a change of course in censorship.
Finally, with regard to crowdsourcing, one should mention the China Digital Times, the University of California at Berkeley project that has crowdsourced URLs and sensitive keywords (Meng, 2011). A well-known story of the use of a sensitive keyword concerns the Chinese term, ‘grass-mud horse’. In Mandarin, this invented term sounds similar to a curse word, and became something of a meme, launching a movement of using terms that are just slightly different from the originals (Yang, 2016). A copious lexicon of these terms was subsequently created, and the China Digital Times project collected them in order to determine the extent to which they are effective in circumventing internet censorship (or ‘censor duping’) in China. Therefore, while the original words or phrases that these terms are similar to may be blocked the question is whether the new ‘grass-mud horse’ sorts of terms are also being blocked. The question leads to further inquiries into the general effectiveness of internet censorship per country and in particular in China – a question formulation discussed below in the case of the effectiveness of Chinese censorship of Falun Gong related websites.

When it comes to researching awareness-raising concerning internet censorship, especially in China, google.com.cn’s redirect to google.com.hk (in 2010) is probably the most significant, having taken place after the public scrutiny in the USA and beyond of Google, Yahoo! and Microsoft for complying with censorship law in China – and also supplying, either voluntarily or through being hacked, information about dissidents to the Chinese
Non-governmental monitoring projects include the ones by Reporters without Borders and Freedom House (with such titles as ‘Enemies of the Internet’ and ‘Internet Freedom’) (Reporters without Borders, 2013). Reporters without Borders, together with Amnesty International, have also performed calendar work and created additional formats to draw attention to the issue. The ‘World Day against Cyber Censorship’ falls annually on 12 March, although it has not recently been as well observed as in the early years after its founding in 2008 (see Figure 4.6). The China Channel Firefox add-on, while no longer functioning, is still worth mentioning when creating a genealogy of internet censorship awareness projects, especially software projects to that end. The extension can still be found, though one probably needs to install a retrograde version of Firefox to use it. The add-on automatically turns on a Chinese proxy in Firefox and gives the user the experience of visiting websites in China. For users outside China it was one of the initial ways to experience Chinese internet censorship (and DNS tampering), whereby the connection one is trying to make to a (forbidden) website times out. Chinese censorship makes it appear that one’s internet service is poor. It also whitewashed the Tiananmen Square massacre, awareness-raising campaigns about which likely contributed

Figure 4.6 The ‘unblocking’ of critical news websites in countries known for internet censorship, on the ‘World Day against Cyber Censorship’, 2016.

Source: Reporters without Borders, 2016a.
to Google's decision to move domains. The Open Net Initiative created a side-by-side browser window experience, with ready-made queries, showing results from the same query in google.com.cn and google.com (Figure 4.5).

Search engine work

The use of search engines to source URLs and make lists seems obvious, and one could consider employing the associational snowball query technique (described in the chapter on issuecrawling) for making a list of a type of website, such as right-wing populist and extremist sites in Germany or critical Iranian bloggers inside and outside Iran. The research question would revolve around whether these websites persist, given their targeting by authorities. Another question would concern whether censorship (or banning) kills content, as remarked above. That is, do the Iranian bloggers or other critical voices continue to blog, even though they are censored in Iran and voided of an audience?

Search engines also may be used to build a country-specific or local list, using advanced search settings, such as setting to Persian, and querying controversial terminology. Here the list of URLs retained would be checked for censorship, with the question concerning the extent to which censors are active in following the use of controversial terminology and blocking sites that use the terms. How active is censorship? Here one relies on Google's algorithm to return significant or relevant sites per language, particularly fresh ones, allowing for some understanding of how quickly censors are on the scene.

Another outcome of search engine work is a national list. A query is formulated in a local domain Google, using a search operator, [site:], and a top-level country domain [site:.uk], and subsequently second-level country domains [site:co.uk]. With the United Kingdom as an example, one builds a list of UK websites by using google.co.uk and the advanced setting, ‘return sites from the UK’. One can query google.co.uk for [site:.uk] as well as a generic keyword, in order to output a list of 1000 websites, or the maximum number of sites returned by Google. One could subsequently lengthen the list of UK websites by querying Google for second-level country domains; rather than querying .uk, one would use .co.uk, .ac.uk, .org.uk and so forth. By following this procedure, one can build a decent list of ranked British URLs, and also subdivide them into lists of companies (.co.uk), universities (.ac.uk), NGOs (.org.uk), and so forth. To build such lists, more generally, a directory of second-level country domains is needed; Wikipedia and Internet registrars are sources for lists of this kind. (The Norwegian internet registrar at norid.no has been known for having one of the more accessible and orderly lists of top-level and second-level country domains.) It is also helpful to know that some countries only have top-level but no second-level country domains, such as the Netherlands (.nl) and Germany (.de). Conversely, France has a wide range of second-level domains, but some are seldom used. Therefore, when conducting preparatory work, one should familiarize oneself not only with which countries have second-level domains but also the national use culture.
Another means of compiling lists of relevant websites per country is the use of Alexa’s top sites by country feature, which is derived from the surfing behaviour of Alexa toolbar users and the postcode they entered when downloading and installing the software. (SimilarWeb also has top 50 country lists.) Here, with Google and Alexa, one is making lists of relevant sites as opposed to lists of categories of sites, in the first instance. ‘Top sites’ from Alexa would be checked for filtering as opposed to category-specific sites such as file-sharing, famous blogger or provocative attire websites, in the site category approach. When one combines the lists from Google site search, the Alexa top sites per country and other devices claiming to output a country or language list of top URLs, the approach could be called (national and language) ‘device studies’.

**Censorship on the Iranian web**

![Diagram](image)

**Figure 4.7** Iranian URL collections, with relative quantities of blockages.  
*Source:* Rogers et al., 2012.

**Device studies**

The fourth technique for building lists is device studies, which largely entails outsourcing the work of list-making to multiple online devices or platforms. In the proof-of-concept project, a study of websites significant to Iranians, to Alexa’s top sites in Iran and Google’s web search results for Iranian websites were added other outputs of devices including blog aggregators, advertising tools as well as social news aggregators and web content rating sites (Rogers et al., 2012). One could consider each distinctive entry point to the web (the searcher’s, the blogger’s, the advertiser’s, etc.), and indeed each outputs relatively distinctive lists of websites (see Figure 4.7). There is a list of sites frequently visited by those based in Iran (Alexa’s list). There is a list of sites in Persian and/or published in Iran which is available in Google’s ‘region’ drop-down menu in the advanced search settings. (There is no google.co.ir, or for that matter local domain Googles for Cuba, North Korea, Sudan or Syria, because of sanctions imposed by the United States Office of Foreign Assets Control that also restrict the use of Google Ads. There is a google.ps for Palestine.) A third kind of device are advertisers’ tools such as Google Display Planner, Adwords or those by other
companies that provide lists of URLs, such as top URLs per country or per language. These are lists of links that are attractive to advertisers for they receive higher amounts of traffic. Blog aggregators are another type of useful device for building lists of URLs. In certain countries, bloggers are considered influential voices (and some are themselves aggregated on ‘global voices’, the Berkman Center project). There may be lists of top bloggers or blogs. In the research on Iran, the bloggers were the most significant targets of the censors, more readily blocked than other types of websites (see Figure 7). Social news aggregators (such as Balatarin in Iran or Reddit) may be used to harvest significant URLs to gain an indication of blogosphere censorship compared to other types (or spheres) of websites (such as top news stories ‘uprated’ by the Iranian crowdsourced aggregator, Balatarin).

As mentioned above, greatfire.org takes a device studies approach, having created lists of websites as well as keywords relevant to China and Chinese users, and charting their blockage (see Figure 4.5). It is an awareness-raising (and transparency) project (mentioned above) that takes a device studies approach, using URL lists sourced from different devices, including Wikipedia, Weibo and Alexa’s list of the top URLs in China, showing levels of censorship per list (see Figure 4.8).

Creating a country-specific URL list with Twitter is possible, though topic-specific lists are more readily prepared from a tweet collection made with multiple keywords and hashtags, whereby the URLs in that collection are extracted. While laborious, for country-specific URL lists one could strive to create a national tweet collection and harvest (and unshorten) referenced URLs, such as in the work on mapping the Australian twittersphere (Bruns et al., 2014).
Dynamic URL sampling (list lengthening)

With the list-building approaches sketched above, sometimes the size of the list of URLs is rather limited. With what is termed ‘dynamic URL sampling’ the lists may be lengthened by crawling websites and capturing outlinks. The crawling technique could be co-link or snowball analysis, and each iteration of the method (one degree, two degrees or three degrees of separation from the seed list) would lengthen the original list. Thus, one may wish to launch multiple crawls, concatenate the lists and remove duplicates, using the triangulation tool. In the sample project the dynamic URL sampling technique is utilized to study the effectiveness of state Chinese censorship of Falun Gong websites.

Measuring the effectiveness of the censorship of the Falun Gong in China

How effective is Chinese state censorship of Falun Gong websites? How ‘well’ does the state censor those websites? Have some slipped through? This project was carried out with proxies, but given the potential in China for next-generation censorship (logging users), it is important to consider using known VPNs or services that monitor Chinese state censorship.

Having curated a list of Falun Gong websites using the techniques discussed above (editorial approach as well as search engine work), and lengthened the list (with the Issuecrawler), one fetches the URLs through the Censorship Explorer, checking for blockage (see Figure 4.9). As mentioned, one should become familiar with the blocking techniques in the country in question through a brief review of the claims (Bamman et al., 2012). In China, note again that the censorship techniques often result in no response code, or a time-out, for the TCP connection may be reset for consecutive requests because (for example) a keyword from a blacklist is in the URL or the site itself is blacklisted. The same effect may be observed if a proxy or VPN is not working. To ensure, first, that the proxy server or VPN is functional, enter a non-censored website and check the response code.

How to be certain, or reasonably so, that a URL is blocked by the state in a certain country (including China)? For national blocking, triangulation or checking multiple ISPs in the country in question is advised, where one seeks agreement between ISP results, while also ensuring that the website is up and running (response code 200) in the Netherlands, for example, which is where the Censorship Explorer tool is based. For selective blocking (second-generation internet censorship), as mentioned above, ISPs from one city (say, the capital or a separatist region) may be targeted rather than those from other locations in the same country. One may consider travelling to a country in question or engaging and informing on-the-ground researchers; however, such work requires knowing the risks.

When using the remote analysis approach – fetching URLs through a proxy, VPN or censorship monitoring service – it is advisable to ascertain information about the services in question, especially concerning whether the server may be traced to an unwitting individual.
Figure 4.9  Issuecrawler input and output. Falun Gong websites inputted into the Issuecrawler, with the resulting map. Use ‘retrieve starting points and network URLs’ for a list of the outputs.
Figure 4.10 Censorship Explorer tool.

(see Figure 4.10). Censorship monitoring services or VPNs designed to deliver sensitive content are preferred, but proxies may be considered in countries not known for logging or expected to log user activity.

In the event, across multiple locations in China, Falun Gong websites are blocked by timing out. The pages do not load after considerable waiting time, making it appear as if they are down, rather than blocked. The censorship is effective across all of the sites checked (see Figure 4.11). It is also effective in sites in both Mandarin as well as English, raising the question of the extent to which blocking is undertaken of sites in even more languages.

Discovering previously unknown censored websites in Iran, and confirming blocking

Around the time of the Iran election crisis (June 2009), considerable efforts were made to study Iranian internet censorship, including the throttling of the connection speeds, the blocking of particular sites as well as (temporary) network outages. In two related projects, researchers and I sought to discover (previously unknown) blocked websites as well as confirm those reported to be blocked. For the confirmation work, a group of regional journalists and I collected press reports of blocked websites in Iran, compiled a list, added to the list URLs from Alexa’s top 20 websites in Iran, and checked them for blocking, thereby both
confirming but also expanding the list of known blocked sites (Caucasian Causes, 2009). So that others could confirm the findings (or at least know whence they came), we also published the proxy IP addresses. At this stage one did not consider whether the research use of another computer user’s proxy, whether it was made available purposively or unknowingly, could put him or her (or the organization) in harm’s way.

For the blocked website discovery, the technique entailed using an expert list of ‘sensitive sites’, crawling them for links to other (related) sites, and fetching the URLs of the newly discovered sites (from the network analysis) in Iran (using proxy servers). Among the findings were that certain newly discovered sensitive websites were still accessible in Iran and that the country also practised page-level (rather than only site-level) blocking as in the case of bbc.co.uk/Persian (Rogers, 2009c). Apart from the proxy issue, here one is confronted with the question of how (and whether) to publish the results (see Figure 4.12). One ‘map’
A Censored Network: Iranian Social, Political and Religious Sites
A hyperlink analysis method for censored Web site discovery.

Figure 4.12 Colour-coded network map of Iranian social, political and religious sites. The hyperlink analysis method for internet censorship research in this case resulted in the discovery of 30 previously unknown blocked sites in Iran. The map, made at the Digital Methods Initiative, appeared in Le Monde on 23 May 2009 (http://www.lemonde.fr/proche-orient/article/2009/05/23/facebook-interdit-d-acces-en-iran-avant-la-presidentielle_1197253_3218.html). It also appeared on the website of the American National Public Radio.

was published in Le Monde (23 May 2009) as well as on the website of the American National Public Radio (NPR), whereupon discussions ensued regarding how the maps are double-edged. The findings are of interest to the censors, too. Future outcomes of such work, including in the subsequent national web study of Iran, would be data-embargoed, and the maps would remain unpublished, or only made available to bona fide researchers in the area of internet censorship research (Rogers et al., 2012).

Conclusions: Contributing to internet censorship research, post Snowden

How to make a contribution to internet censorship research, while considering whether it is conceivable that the researcher and/or the computer users in the country in question will be placed in harm’s way? Covert research continues but, given the risks, it is challenging
to repeat and scale. Remote research still may be practised, but much caution should be taken. Since the Snowden revelations of widespread state surveillance, and the publication of techniques via WikiLeaks, research has concentrated on detecting whether Western filtering and monitoring software is in use in authoritarian countries. In-country software, however, may have been developed to monitor and filter, too.

The notions of ‘network authoritarianism’ and third-generation internet censorship have been put forward to capture how states achieve censorship through monitoring in the first place, and secondarily through filtering (MacKinnon, 2011). That is, while there is blocking, a state follows the online chatter, and intervenes in the content spaces with correctives as well as noise. Following the corrective information, disinformation, ‘fake news’ and other content interventions become a means to study not so much censorship but rather content control, reach and spread. The effects of network authoritarianism become the objects of study rather than the logging and interpretation of blocked URLs and keywords.

Among the work to be done is the development of techniques to discover networks of content interventions and their effects (see Chapter 11 on tracking for network discovery and content interpretation techniques). The research questions developed for internet censorship are still of relevance. Slightly rephrased, they may be put to use to study such questions as content network discovery, unintended consequences, effectiveness, circumvention, monitor duping, Western complicity and monitored content effects.
PROJECT 3

Investigate the effectiveness of state internet censorship in China

RESEARCH GOAL To determine the effectiveness of Chinese state internet censorship by building and analysing a list of websites the authorities are known to block.

1. Build a list of Falun Gong websites through an editorial approach or associative query snowballing (search engine work).
2. Optional. Crawl and expand the list (of related sites) with the Issuecrawler (dynamic URL sampling). Extract the expanded list of websites from the Issuecrawler (using ‘retrieve starting points and network URLs’ feature). Remove off-topic URLs.
3. Run the list of URLs (at least hosts) through the Censorship Explorer software, choosing a variety of ISPs, or by other means, considering the ethical issues discussed above. One could consider using ISPs that are geographically dispersed in order to check for regional differences in filtering (such as in Macau and Hong Kong). One also may wish to spot-check using the Blocked in China website (blockedinchina.net or the test URL feature at greatfire.org).
4. If using the Issuecrawler, colour-code (or greyscale-code) the map (manually), using Adobe Illustrator or similar, indicating blocked and unblocked sites.
5. Note the effectiveness of the censorship across the landscape of Falun Gong sites.
6. As an additional undertaking, consider curating lists of Falun Gong sites (or of other sensitive subject matters) in additional languages, and checking for blockages, in order to ascertain the reach of the effectiveness of state internet censorship.

PROJECT 4

Investigate state internet censorship in Iran through a discovery technique

RESEARCH GOAL To discover previously unknown censored websites in Iran using the Censorship Explorer and the Issuecrawler.

1. Obtain or build a list of websites known to be blocked in Iran. As discussed above, non-governmental organizations and Wikipedia articles have descriptions of the (types of) sites blocked or lists thereof. One may also wish to compile lists from news reports and other sources.
2. Optional crawl and expand the list (of related sites) with the Issuecrawler (dynamic URL sampling or another hyperlink analysis tool).
Check the (expanded) URL list for blocking, through the Censorship Explorer or another means, considering the ethical issues discussed above.

Colour-code the map, or annotate a list (after removing off-topic URLs), with three types of nodes: blocked, unblocked and newly discovered blocked.

Are there newly discovered blocked websites on the map (or on the list)? Consider adding them to the record of known blocked websites (on Wikipedia or elsewhere). Are there sensitive websites on the map (or on the list) that are not (yet) blocked? Consider how to report this information yet embargo the details.

---

**Project 5**

**Conduct internet censorship research in one or more countries**

**Research Goal** To discover or confirm and report on online content blocking in a country or countries of choice.

1. Choose one or more countries, and build a list of websites to be checked for blocking.
   - The amount of internet censorship research available on that country. What is known, and what is claimed? For internet censorship research per country and region as well as for claims, see (among other sources) http://www.opennet.net/, http://www.freedomhouse.org/ and http://www.rsf.org/.
   - General contribution to make. Where are the overall gaps in knowledge about censorship? Are there urgent analytical needs, such as new claims about censorship or current events which could give rise to blocking, including upcoming elections or planned protests?
   - Research questions to be posed. Does the research concern discovery, unintended consequences, effectiveness, circumvention, censor duping and effects on the content being censored or its authors being watched (self-censorship)? Is it striving to do detective work about the filtering software being used (and who supplies it, such as Western software firms)?
   - URL list-building. Consider techniques such as editorial, search engine work and so forth as well as types of websites to be checked, such as categories of sites (OpenNet-style or other), top sites per country (e.g., Alexa, Google queries), platforms (e.g., social networking sites), event-based (e.g., elections) and so forth. Consider also the lengthiness of the list of (types of) sites suspected of being blocked in order to make general claims. Shorter lists may be as powerful as longer ones, if they are the ‘top sites’. Make a case for your list-making technique that is a reasoned consideration of whether your list allows you to test claims.

2. Check for blocking: consider VPNs, proxies and verification strategies. If in existence in the country in question, consider using an existing censorship monitoring service such as greatfire.org for China.
   - The Censorship Explorer allows you to check URLs for blocking via VPN or proxies.
CONDUCT INTERNET CENSORSHIP RESEARCH IN ONE OR MORE COUNTRIES

- If using a VPN, install the DMI browser toolbar, https://wiki.digitalmethods.net/Dmi/FirefoxToolBar. The Censorship Explorer opens the URL via the DMI toolbar in the browser and sends the result to the Censorship Explorer. If using a VPN on your computer, the URL is retrieved via your VPN.
- The Censorship Explorer provides a list of open HTTP proxies, often scraped from other sites. You can also add other proxies (e.g., found by searching for [free proxy lists]). Although all proxies listed in the Censorship Explorer are publicly available on the web, they might not all be made accessible intentionally. Open proxies can mistakenly be left open and found by sites scanning for open proxies. Consider the political situation in the country of the proxy you are using in relation to the types of URLs you wish to fetch through the proxy. URL blocking is mostly intentional and web traffic from a proxy to a blocked site may have consequences for the owner of the proxy. For ethical reasons, use VPNs (check their terms of service to see whether you can request certain types of content via their service) or institutional proxies (such as from universities) so as not to put individuals in harm’s way. Avoid having friends or acquaintances check URLs on site in the countries where censorship is practised.
- How to build trust in the results returned by the VPN or proxy? For open proxies it is often challenging to infer the owner, its location and its representativeness for a country, region or city. Before checking for URL blocking, endeavour to find out who owns the VPN or proxy and where it is located. Thereafter test the VPN or proxies with two unblocked URLs to ensure they are working. Use proxies that produce consistent and reliable results.
- How do you know that a URL is blocked? The Censorship Explorer automatically also queries each URL from a Dutch IP address, in order to verify VPN or proxy results with control results from a country that does not censor. Blocking can transpire in different ways. Do the response codes differ when accessed through the VPN or proxy versus when accessed through the control server? That is, does the control server produce consistent 200 OK HTTP status codes, and the VPN or proxy output different results? Does the connection time out? Is there a redirect to an unexpected page?
- What is the type of blocking? Is it transparent (e.g., does it provide a block page)?
- What type of proxy was used? Does that influence the results? Note that transparent proxies do not modify the request or response beyond what is required for proxy authentication and identification, and non-transparent proxies modify the request or response in order to provide some added service to the user agent, such as anonymity filtering.
- Where is the VPN or proxy located? On which network (ISP) does the proxy reside? Do you receive different results for different proxies in the same country? To obtain answers to these questions you can use tools such as http://whatismyipaddress.com/ip-lookup or http://www.whatismyip.com/tools/ip-address-lookup.asp.

3 Document findings.
- Make a list of the tests in a spreadsheet. For each URL tested and each proxy used, the list should at least contain the IP address and port of the proxy, date and time of the test, response code returned, type of proxy used, the network (ISP) on which the proxy resides and the geographical location of the proxy.

4 Discuss implications.
- Discussion of the findings, and contribution to internet censorship research, in light of the literature and current state of the field. How does the research contribute to discovery, unintended consequences, effectiveness, circumvention, censor duping, effects on the content being censored, self-censorship, or (Western) filtering software detection?
• Is the contribution newsworthy?
• Consider introducing both the limitations of the work as well as the needs for future research, be it conceptual or methodological.

Tools

- Issuecrawler, issuecrawler.net. Available at https://www.issuecrawler.net.
- Triangulation, digitalmethods.net. Available at https://wiki.digitalmethods.net/Dmi/ToolTriangulation.
- Test if a site is blocked in China, https://www.comparitech.com/privacy-security-tools/blockedinchina/

Resources by list-building approach

Editorial approach

- Open Net Initiative, http://opennet.net/
- Blocked on Weibo, http://blockedonweibo.tumblr.com/

Crowdsourcing

- Amnesty International’s anniversary project, http://irrepressible.info/ (look up in the Wayback Machine at archive.org)
- Great Firewall of China, GreatFire.org, https://greatfire.org/

Search engine work

- Google search operators, https://support.google.com/websearch/answer/2466433

Device studies

- Online censorship in China, GreatFire.org, https://greatfire.org/
- SimilarWeb, Top 50 sites by country, https://www.similarweb.com/top-websites

Dynamic URL sampling

- Issuecrawler scenarios of use, http://www.govcom.org/scenarios_use.htm
- Issuecrawler FAQ, https://wiki.issuecrawler.net/Issuecrawler/FAQ