
Research Publication No. 2003-02
4/2003

# Empirical Analysis of Internet Filtering in China

Jonathan Zittrain
Benjamin Edelman

# Empirical Analysis of Internet Filtering in China[†]

**Jonathan Zittrain and Benjamin Edelman**
Berkman Center for Internet & Society, Harvard Law School

The government of the People's Republic of China has a longstanding set of policies restricting their citizens' exposure to information. The Internet poses a new challenge to such censorship because of the breadth of online content, the rapidity with which sources of content can be moved or mirrored, and because content sources are often remote from Chinese jurisdiction.

A desire to capture the economic benefits of networked computing while maintaining control over the public's Internet exposure has led to a variety of strategies to split the difference between allowing unfettered access to the global Net and refusing to countenance any deployment beyond trusted elites.

We collected data on the methods, scope, and depth of selective barriers to Internet usage through networks in China. Tests conducted from May through November 2002 indicated at least four distinct and independently operable Internet filtering methods — web server IP address, DNS server IP address, keyword, and DNS redirection — with a quantifiable leap in filtering sophistication beginning in September 2002.

As with most technical filtering regimes, whether implemented at the client, Internet service provider, or backbone level, no list of the sites blocked or the methodologies used to block them has been made available by those doing the filtering. Further, while the government-connected Internet Society of China (not a chapter of the well-known non-profit international Internet Society, <http://www.isoc.org>) has asked ISPs and content creators to sign a pledge that includes self-filtering, few official statements document that government-maintained Web filtering exists, much

---

[†] This article is an update and revision to work originally posted to <http://cyber.law.harvard.edu/filtering/china/> and subsequently published as "Internet Filtering in China" in the March-April 2003 issue of IEEE Internet Computing.

less the criteria employed and thresholds necessary to elicit a block. We therefore investigated the growing methods of Internet filtering, and collected and distributed a list of blocked sites and pages on our web site at <http://cyber.law.harvard.edu/filtering/china> — a diverse list that is large in absolute terms, even if small relative to the size of the Internet and to the total amount of still-undocumented blocked content. Such a list lets us assess the nature and scope of filtering in China, paying particular attention to inaccessible non-sexually-explicit Web sites.

## Testing Methods

Our testing relied on two separate data collection methods — through modems and open proxy servers. From 20 March to 6 May 2002, we connected with an international telephone call by modem using dialup accounts with several Chinese ISPs. After 6 May, our modems were unable to negotiate a handshake with modems answering at any Chinese ISPs, a failure consistent across multiple phone lines, locations, multiple ISPs, and points of presence in China. From 14 August to 12 November 2002 we connected to open proxy servers in China. We selected open proxies and determined their listed locations for tabulation purposes using APNIC's IP-WHOIS.

During testing, we requested 204,012 distinct sites drawn from various Web indices (such as sites listed in Yahoo Taiwan's (tw.yahoo.com) directory categories and Yahoo's Taiwan subdirectory categories (<http://dir.yahoo.com/Regional/Countries/Taiwan>)), and search results (such as search engine google.com's top 100 results for a search on "China freedom"). Most sites were accessible from China just as from our standard Internet connection in the United States, but we found that certain sites were consistently unavailable. By attempting to retrieve these sites repeatedly over time, from multiple locations in China, we drew inferences on which specific sites among them were intentionally blocked by Chinese network staff. In this way, we found that 18,931 sites were inaccessible from at least two distinct proxy servers within China on at least two

distinct days while still accessible from the US. The sites we tested were in no way intended to be "representative" of the World Wide Web; instead, we tested sets of sites that might be selected for blocking in order to generate as lengthy a list of blocked sites as possible. In some instances it may be difficult to distinguish between intentional blocking and unintended network glitches; based on the number of times that each page was accessible versus inaccessible, the data appendices that accompany the online version of this article attempt to indicate our relative certainty as to blocking of each listed site.

## The Scope of Filtering

We tested one URL per Web host – the "default," i.e. "front page" URL – based on reports, confirmed in subsequent testing, that when the default page of a site was filtered the entirety of that site was typically filtered. As a result, when we report a site as inaccessible, the entire site was generally inaccessible — not just the site's default page or front page.

To test the hypothesis of entire-site blocking, we formed a sample of inaccessible Web hosts and checked whether an arbitrary subdirectory on each such site was inaccessible. Though the arbitrary directory name we chose was intended not to exist on the servers, typical Web servers return a "not found" error message in response to a non-existent request. These error pages themselves were inaccessible in 99.8 percent of the tests. We attribute the other 0.2 percent to anomalies (such as transient network errors that might have wrongly rendered the Web host inaccessible in the first instance when the host was not intentionally blocked).

At the moment, then, it seems that when the host default page is blocked, all other pages on that host are also blocked. Of course, the reverse need not be the case, and we have separately confirmed multiple instances in which it is not the case. For example, China has blocked access to <http://cyber.law.harvard.edu/filtering/china>, the Web site that contains much of our recent

writing about China's filtering efforts. However, the rest of the cyber.law.harvard.edu Web server remains accessible. Thus, at least some blocking appears to be triggered by relatively few keywords in page URLs or contents, representing a technical layer of blocking wholly distinct from (and seemingly rarer than) an entire site being made unavailable. See the "Filtering Implementations" appendix for more information, including summaries of the newer DNS server IP address, keyword, and DNS redirection methods of Chinese Internet filtering.

When an entire Web host is filtered, our data shows that this filtering typically operates on the basis of the host's IP addresses rather than on one or several domain names. To confirm this, we observed that when distinct Web sites are hosted on a single Web server (as is typical in commercial "shared hosting" at the lowest monthly rates), blocking one Web site on a given server (with a given IP address) requires blocking all Web sites on that server. For example, we found 308 distinct blocked sites (by domain name and differing page content) all hosted on the server at IP address 216.34.94.186, a parking/redirection server used by domain name registrar Dotster. This server hosts additional Web sites beyond those we tested, and it is highly likely that they too were blocked. Indeed, a representative from domain name registrar Enom reported that its primary domain name forwarding service had been blocked by China — rendering literally thousands of domain names unreachable. In subsequent work, <http://cyber.law.harvard.edu/people/edelman/ip-sharing>, Edelman has found that more than 87% of .COM, .NET, and .ORG domain names share their web server IP addresses with one or more other domains, and two thirds of domains share their web server with fifty or more others. These results suggest that China's IP-based filtering systems may be responsible for much of the blocking we have observed of content that to us seems unobjectionable.

## Sexually Explicit Content Filtering

A preliminary round of testing examined 795 distinct URLs containing sexually explicit images. These URLs had been used as the basis for a portion of Benjamin Edelman's expert testimony in *Multnomah County Public Library, et al. v. United States* (<http://cyber.law.harvard.edu/people/edelman/mul-v-us/>). He generated this list by collecting all 797 results from Google listings in response to an October 2001 Web search using the search criteria "free adult sex." He removed two pages because they didn't include sexually explicit images. Of the 752 pages still providing content at the time of our testing, 101 were blocked in China (13.4%). Edelman previously found that leading commercial filtering applications blocked 70 percent to 90 percent of these sites (<http://cyber.law.harvard.edu/people/edelman/pubs/aclu-113001.pdf>). We infer from this that China (unlike Saudi Arabia, given data at <http://cyber.law.harvard.edu/filtering/saudiarabia>) has not relied upon commercial filtering applications to salt its own list of blocked sites of this sort.

## Non-Sexually Explicit Content Filtering

Our main testing examined Web sites drawn from categories other than sexually explicit content. We seeded this site list from multiple sources. For example, we extracted from Yahoo all Web sites in certain categories (including those specifically about education, entertainment, news, major world governments, and politics) as well as all sites in the non-English regional versions of Yahoo that specifically concern China and Taiwan (cn.dir.yahoo.com and tw.dir.yahoo.com). We conducted searches on terms likely to yield sensitive results and thus candidates for blocking, both in English and in Chinese, using the Google search engine, and placed the top results into our list of URLs to test. We tracked approximately 5,000 additional sites submitted by Internet users to our Real-Time Testing System (<http://cyber.law.harvard.edu/filtering/china/test>) through September 2002, and we received email suggestions of further sites to test. The result of these data sources

was a list of 203,217 distinct host names.

We found that a total of 18,931 of these sites (9.3 percent) were blocked in China. A full listing of blocked sites is available at <http://cyber.law.harvard.edu/filtering/china>.

### Content Not Filtered

Many sites are not blocked in China, whether because they have yet to be passed upon by the authorities that determine blocks or because they have been affirmatively found to be nonsensitive. Sites not blocked might assist in drawing inferences about what content among the blocked sites is responsible for the differential treatment, or how assiduously a given objection to certain types of content is enforced. For example, filtering of the official site for the United States Federal courts (uscourts.gov and all subdomains) might indicate a desire to prevent access to information about the American judicial system, its processes, and its rulings — but Findlaw, LexisNexis, and Westlaw all remain accessible. Similarly, blocking of well-known sexually explicit sites such as playboy.com and penthouse.com suggests a purposeful decision to restrict sexually-explicit material — yet hustler.com and whitehouse.com were consistently accessible in our testing.

## A Taxonomy of Blocked Sites

Our online report provides a full listing of some 19,000+ specific web sites found to be inaccessible from China. A full print listing of these many URLs is beyond the scope of this article, but we report below a general taxonomy of blocked sites.

We found that blocking varied across different proxies in China, reinforcing the notion that blocking is not done through a central bottleneck. However, there is insufficient data to draw conclusions about systematic variations in blocking across geographic locations; current data is consistent both with intentional variations in blocking and with delays in updating block lists in certain regions.

We obtained selected sites from Google searches on designated keywords. Figure 1 shows a sampling of the sites blocked in response to searches on specific keywords.

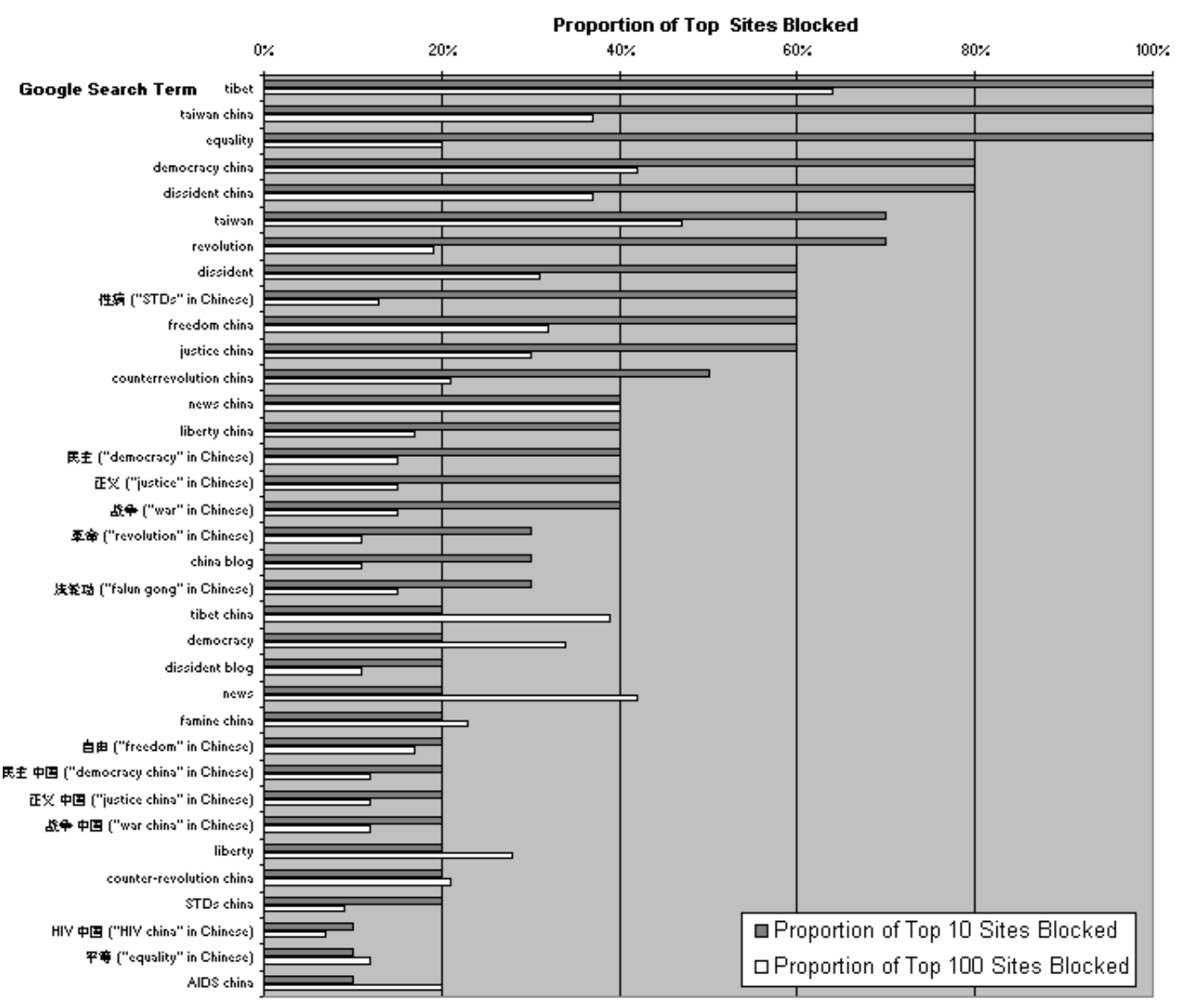


*Figure 1. Proportion of sites blocked by Google search term. This figure reports the proportion of sites blocked, among sites suggested by Google in response to searches on particular keywords.*

## Dissident/Democracy Sites

Of the top 100 sites Google returned for a "democracy china" search, 40 were blocked, while 37 "dissident china" sites were blocked, 32 were blocked for "freedom china," and 30 for "justice china." Specific blocked sites included Amnesty International, Human Rights Watch, the

Hong Kong Voice of Democracy, the Direct Democracy Center, and dozens of Falun Gong and Falun Dafa sites.

### Health

Of the top 100 Google results for "hunger china," 24 were blocked; 23 for "famine china;" 21 for "AIDS china," 19 for "sex china," and 14 for "disease china." Specific blocked sites included the AIDS Healthcare Foundation, the Internet Mental Health reference, and the Health in China research project. We found 139 sites listed in Yahoo's health directory categories and subcategories blocked.

### Education

Several well-known institutions of higher education, including the primary Web servers operated by Caltech, Columbia, MIT, and the University of Virginia were blocked. Non-university sites including the Learning Channel, the Islamic Virtual School, the Music Academy of Zheng, and the Web sites of dozens of public and private primary and secondary schools were also blocked. We further found evidence of blocking of 696 sites listed in Yahoo's education directory categories and subcategories.

### News

The BBC News was consistently unreachable. CNN, *Time Magazine*, PBS, the *Miami Herald*, and the *Philadelphia Inquirer* Web sites were often unavailable as well. Of Google's top 100 results for news, 42 were blocked. We further found evidence of 923 blocked sites listed in Yahoo's news and media directory categories and subcategories. Nonetheless, some news sites that were previously blocked became accessible during the course of our testing. For example, Reuters was blocked through 29 April, but was subsequently accessible, while the Washington Post was blocked through 6 May and was subsequently accessible. This reduction in blocking of entire news

sites might reflect that certain new filtering technologies (including keyword-based filtering) that allow blocking of particular sections and articles that are particularly controversial in China. Thus, our results should not be taken to suggest that every Washington Post article is accessible in China, even when the IP address of washingtonpost.com is not blocked.

### Government Sites

Blocked sites included a variety of those operated by governments in Asia and beyond. The entire site of uscourts.gov, including the many federal district and appellate courts in the United States, as well as the United Kingdom's Court Service and Israel's Judicial Authority were blocked. The communication sites of various governments were blocked, including the U.S. government's Voice of America, as well as travel sites for Australia, Israel, Korea, Switzerland, and Wales. Government military department sites were also blocked, including the U.S. Department of Defense, though others remained reachable (such as the CIA). A variety of additional government sites were blocked, without manifest pattern, both in the U.S. and beyond. Examples include the site of Seattle's King County, the main Australian Federal Government index site, the Philippines Bureau of Customs, the British Insolvency Service, the Office of the Governor of Makkah in Saudi Arabia, and the Legislative Assembly of British Columbia. Blocked sites included 516 of those in Yahoo's categories and subcategories pertaining to governments.

### Taiwanese and Tibetan Sites

Blocked sites included business sites (like the A&D Company of Taiwan), noncommercial sites (the Taiwan Health Clinic and a total of 709 .edu.tw sites, as well as the Voice of Tibet), and government sites (the Office of the President of Taiwan and the Taiwanese Parliamentary Library among 936 other Taiwanese government sites, and the official Web site of "the Tibetan Government in Exile"). More than 60 percent of Google's top 100 Tibet sites were blocked, as well

as more than 47 percent of the top Taiwan sites. Taiwanese content was also blocked disproportionately, relative to its representation in our testing sample; fully 3,284 .TW sites (13.4 percent of .TW sites tested) were blocked, while our overall block rate was approximately 9.3 percent. (Of course, comparisons of block rates must be performed with care given the subjective formation of the list of sites tested. For lack of a domain name specifically associated with Tibetan sites, it is more difficult to perform such a comparison on the block rate of Tibetan content.)

### Entertainment

Blocked sites included the movie Deep Impact, the Canadian Music Centre, the Taiwanese site of MTV (mtv.com.tw) and multiple sites providing off-color jokes. We also found blocking of a total of 451 sites in Yahoo's categories and subcategories pertaining to entertainment.

### Religion

Blocked sites included the Asian-American Baptist Church, the Atheist Network, the Catholic Civil Rights League, Feng Shui at Geomancy.net, the Canberra Islamic Centre, the Jewish Federation of Winnipeg, and the Denver Zen Center. We found 1,763 sites in Yahoo's categories and subcategories pertaining to religion blocked.

## Conclusions

From our data over time, it appears that the set of sites blocked in China is by no means static. Whoever maintains the block lists is actively updating them, giving special attention to certain general-interest high-profile sites where content changes frequently. This is particularly noticeable with news sites such as CNN and Slashdot. Some new sites with sensitive content are promptly blocked. However, even some longstanding sites of apparent sensitivity remain unblocked. This is most easily noticed in our data for sexually-explicit sites — we found blocking of only 13.4 percent of our sample of well-known sexually-explicit sites — but it is also

anecdotally apparent from our finding that, for example, some U.S. intelligence sites are blocked while others are accessible. Further data collection will be geared toward determining the extent to which the basket of sites blocked reflects shifting substantive government policies — whether, for example, a change in relations with Taiwan is reflected in blocking, and if so, how quickly. Meanwhile, experience of the past year suggests that filtering of western news sites – as well as search engines like Google and other sources potentially critical of the Chinese government – seems to tighten in the weeks before and after key political events such as the March 2003 Party Congress.

China's Internet filtering efforts remain opaque, and in the absence of government cooperation or admission of filtering methods, data probing of the sort used in our study is intended to help determine the scope of filtering. We have previously studied filtering in Saudi Arabia and in American public libraries (<http://cyber.law.harvard.edu/filtering/saudiarabia> and <http://cyber.law.harvard.edu/people/edelman/mul-v-us>); in these locations, blockage of a Web page leads to an error message clearly explaining that the requested page is unavailable due to intentional blockage. In contrast, China's systems make it difficult for a user to distinguish between an intentional block and a temporary network or server glitch. This might be by design or might reflect technical happenstance — that this implementation was easier or cheaper, given the size and design of China's network infrastructure. But some newer forms of Chinese filtering — namely, redirection of a request for a sensitive Web site to another Web site — can be either more or less obvious to the user than an apparent network glitch, depending on whether the substitution is noticed.

The Chinese government and associated network authorities are clearly continuing to experiment with different forms of blocking, indicating that Chinese network filtering is an

important instrument of state Internet policy, and one to which significant technical and human resources continue to be devoted.

## Acknowledgments

The authors are grateful to Ronald Guilmette for assistance with locating proxy servers in China, to Joshua Rosenzweig of the Dui Hua Foundation for assistance in locating routing glitches, and to Nongji Zhang of the Harvard Law School Library for assistance with Chinese translations.

An electronic version of this document, with full data appendices, is available at <http://cyber.law.harvard.edu/filtering/china>.

**Jonathan Zittrain** is the Jack N. and Lillian R. Berkman Assistant Professor of Entrepreneurial Legal Studies at the Harvard Law School. zittrain@law.harvard.edu, <http://www.jz.org>.

**Benjamin Edelman** is a student at the Harvard Law School and a fellow at its Berkman Center for Internet & Society. edelman@law.harvard.edu, <http://cyber.harvard.edu/edelman>.

## Appendix: Filtering Implementations

On the basis of our testing, both automated and manual, we have reached an increased understanding of the design of filtering systems used to restrict Internet access in China. We have observed certain idiosyncrasies in Chinese methods of Internet filtering, and in some instances we have found methods to circumvent particular aspects of filtering. Based on this data, we can draw inferences about particular methods of filtering.

### Web Server IP Address

We confirmed that filtering operates on the basis of IP address by observing that when China blocked access to one Web site on a given physical server, all other sites on that physical server (that is, on that IP address) were also typically blocked.

Our data suggest that when Chinese network staff deem a site to contain undesirable content, their most common method of filtering it is simply to drop IP packets destined for it. This method likely relies on block lists loaded into border routers that connect China's internal networks with international networks. ISPs reportedly share block lists, perhaps with additional centralized coordination of updates. Variation across networks and over time is to be expected based on delays in propagating list revisions. As a result of these delays and variations, it is often difficult to conclude that a site is "blocked in China," for a given site might truly be reachable from some parts of China and blocked from others.

This method of blocking, the most widely used in our experience, is difficult to circumvent. The typical circumvention method relies on channeling Web page requests and viewing associated results through proxy servers or virtual private networks located outside China. As others have noted,[1] however, monitoring and proxy-blocking efforts provide a check on the use of proxies. When Google's cache feature was available in China, it allowed users to circumvent this method of

filtering, but this feature has since become unavailable due to more selective Chinese filtering of Google use, even as google.com itself is, at the moment, accessible.

### Domain Name Server IP Address

Like filtering on the basis of Web server IP address, this method likely relies on block lists loaded into border routers. Even if the desired Web server is reachable, a user's computer cannot reach the Web server if it cannot first convert the server's domain name into a numeric IP address — and when the site's DNS server is blocked, no such conversion is possible.

We have observed that many of the filtered DNS servers are also themselves Web servers, or are located on networks that are filtered in totality (as distinguished from partially filtered networks for which certain IP addresses are filtered while others remain accessible). This lends some support to the inference that DNS-level filtering might be unintentional — an accidental consequence of filtering a Web server or network that also happens to offer domain name services.

When filtering operates on the basis of DNS IP address, users can sometimes circumvent it by directly entering the desired Web server's IP address. In particular, an interested user might simply enter the IP address of the desired Web server directly into a browser's location bar (into the same location where the site's domain name would ordinarily be placed). Of course, this method requires that the user know the server's IP address, and it further requires that the server provide only this single site (rather than hosting many sites via HTTP multiplexing). Nonetheless, in some situations entering an IP address directly might circumvent Chinese filtering efforts. Another possible circumvention method is the use of non-Chinese DNS servers, with such servers performing a subset of the role that an overseas proxy would serve to circumvent Web host IP blocking. If such an approach became widespread, border routers could be reconfigured to refuse outbound DNS requests except when received from authorized DNS servers.

## DNS Redirection

DNS servers in China have been found to offer seemingly intentionally incorrect answers to the IP addresses of certain domain names. For 1,043 tested sites, we confirmed that DNS servers in China report a Web server other than the official Web server actually designated via each site's authoritative name servers. We call this phenomenon "DNS redirection," though others sometimes refer to the situation as "DNS hijacking." Consistent with prior reporting (www.dit-inc.us/report/hj.htm), our data show that such sites were consistently unreachable in their entirety.

When a user in China requests a site affected by DNS redirection, for example, the user's computer is told that the site's domain name is associated with the IP address 64.33.88.161. That IP address is associated with the host www.falundafa.ca, the site of a Canadian organization that promotes the practice of Falun Gong. However, that address is blocked by Chinese border routers, preventing such requests from reaching either the falundafa server or any other.

While we cannot know for sure the specific rationale for implementing this additional filtering method, we suggest two possible understandings. First, this filtering method might be intended to supplement border router filtering. Depending on the specific implementation method, it might be somewhat more efficient or easily updated by Chinese network staff, and ISP compliance can be more easily monitored remotely via ordinary DNS tools such as "dig." Second, this filtering method is a likely precursor to efforts both to monitor access to specific sites and to revise or replace content on those sites with other content specifically provided by Chinese network staff. Either approach would rely on proxy servers placed at specified IP addresses and would require that requests for designated sites in some way be redirected to those addresses. While this second theory is largely speculative, it dovetails with the Chinese efforts we have documented to replace (and not simply block) Google, and subsequent filtering of certain Google search terms (including the names of key political figures and the terms required to use the Google cache).

Use of non-Chinese DNS servers bypasses this filtering method, though future use might be blocked by border routers.

### URL Keyword Filtering

Beginning in September 2002, our data reflects that a subscriber to a Chinese ISP would receive no response when seeking a URL that contained certain words or phrases. This effect was particularly notable at Google, where names of key political figures are apparently off-limits, as are certain other words used to invoke particular Google features (among them the caching feature that can provide a method of circumventing the filtering implementations described above). In some instances, we have also observed that these keyword blocks apply equally to requests from other sites. From at least certain locations in China, attempts to retrieve any URL containing the character string "jiang+zemin" triggers a distinct kind of temporary filtering (even if the result of that request would only be a "404 -- Not Found" error page).

Subsequent to a request for a URL with a prohibited term, we have confirmed "timeout" periods of 5 to 30 minutes during which either the target site or even all sites (including otherwise-unfiltered sites) became inaccessible. We have received further reports that some timeout periods can last until a user's computer is rebooted or until a user's DSL modem is powercycled. If intentional, as seems likely, this represents a type of filtering that tries to "train" the end user to avoid using prohibited terms, imposing a penalty beyond mere inaccessibility of the requested URL should the terms be used.

This method of filtering is likely implemented via packet-filtering systems integrated into border routers or placed adjacent to them. We have observed that keyword-based filtering systems tend to search for plaintext in URL strings — searching for the word "cache," for example, and blocking any request to Google that contains this word in its URL. However, the HTTP RFC

specification describes additional techniques for encoding ("escaping") characters in a URL.[2] For example, plain text characters can be encoded via escape sequences of the form %4A where 4A is the hexadecimal code of the ASCII character at issue. We have confirmed that in at least some instances, Chinese filtering systems of this sort are not currently triggered by escape-sequenced keywords that, when expressed in plain text, consistently prevent access to the requested pages. (This errata reflects a failure to properly implement the comparison specified in RFC 2616 section 3.2.3.)

Experience in other contexts suggests that packet-filtering can cause performance problems, for its inspection of the content of TCP/IP packets is often slower than the maximum speed at which routers and international lines could otherwise pass packets. Data suggests that Chinese filtering administrators have addressed this problem in at least three ways: 1) By allowing "overflow" packets to pass without packet-filtering inspection or blockage, causing occasional access to web sites and pages that are otherwise accessible. 2) Configuring differential routing so that packet filters need only inspect transmissions to and from hosts of particular concern. 3) Reducing the throughput of high-speed links and allowing performance to suffer. In the long run, the authors believe that the speed of filtering systems is likely to increase more rapidly than the speed of international data communications, suggesting that technical advances may relax these tradeoffs in the future. However, the use of ever more sophisticated filtering rules – requiring the comparison of each packet with an increasingly lengthy list of filtering criteria – may cause continued performance problems.

### Keyword Filtering Based on HTML Response

Beginning in September 2002, we observed that certain keywords within Web page data being transmitted to a Chinese Internet user triggered filtering of that data. In particular, even when

a page came from a server not otherwise filtered, and even when the page featured a URL without controversial search terms, it might nonetheless be inaccessible if the page itself contained particular controversial terms. Such pages were often – but not always – truncated, that is, interrupted midway through their display. On certain browsers, including recent versions of Microsoft Internet Explorer, pages truncated in this way might flash briefly on screen, then disappear. This phenomenon represents an augmentation of "compiled" filtering with "interpreted" filtering — the former representing specific sites deemed *ex ante* to be off limits, with routers configured accordingly, and the latter representing data deemed on-the-fly (mechanically), to be off limits, with corresponding temporary loss of access to the source of that data.

The observed results are precisely what would be expected if Chinese border routers (or associated hardware) implemented a packet-filtering system triggered by particular controversial keywords. To reduce memory and processor requirements, such systems promptly pass on all packets found to be acceptable. However, upon receiving the first packet containing a prohibited term, a packet-filtering system would be configured to discard all further packets from the same source and destination for some designated period — causing the page truncation consistently observed under these circumstances. Observed disuniformity of filtering might reflect that packet filtering operates at less than line speed, that is, it can inspect only a portion of content passing through a given router, as described above. It might also reflect that packet filtering fails to take account of borders between packets, such that a page is permitted to be viewed if a part of a prohibited word is received in one packet and the remainder in a subsequent packet.

Based on our understanding of the likely implementation method of such filtering, we note two possible means of circumventing this filtering. First, content providers can escape their text, using HTML markup that is equivalent to the characters at issue or adding HTML whitespace

(comment tags, and so on) in the middle of controversial words or phrases. (These techniques are documented in HTML specifications for character entity references and comments.) Second, Chinese users can reduce their TCP/IP stack's specified maximum transmission unit (MTU) — reducing the amount of text contained in a given packet and thereby reducing the effectiveness of packet-inspection systems; however, this approach typically reduces performance and also increases network overhead.

The performance problems flagged in "URL Keyword Filtering," above, apply equally to keyword filtering based on HTML response.

In future work, we will seek to document the specific keywords found to be prohibited in searches, URLs, and HTML response pages, and more important, the evolving prevalence of each type of filtering.

## References


1. B. Haselton, "List of Possible Weaknesses in Systems to Circumvent Internet Censorship," November 2002; www.peacefire.org/circumventor/list-of-possible-weaknesses.html

2. T. Berners-Lee, "Uniform Resource Identifiers (URI): Generic Syntax," IETF RFC 2396, August 1998; www.ietf.org/rfc/rfc2396.txt.

# Appendix: Related Web Resources

- Empirical Analysis of Internet Filtering in China (electronic version with full listing of specific blocked sites): http://cyber.law.harvard.edu/filtering/china
- Real-Time Testing of Internet Filtering in China: http://cyber.law.harvard.edu/filtering/china/test
- Web Sites Sharing IP Addresses: Prevalence and Significance: http://cyber.law.harvard.edu/people/edelman/ip-sharing/
- Replacement of Google with Alternative Search Systems in China: http://cyber.law.harvard.edu/filtering/china/google-replacements
- Internet Filtering in the Kingdom of Saudi Arabia: http://cyber.law.harvard.edu/filtering/saudiarabia
- Documentation of Internet Filtering Worldwide: http://cyber.law.harvard.edu/filtering
- Forbidden Sites Hijacked All Over China: http://www.dit-inc.us/report/hj.htm