Citizen Lab Occasional Paper #1

Search Monitor Project: Toward a Measure of Transparency

Nart Villeneuve[1]

Abstract This report interrogates and compares the censorship practices of the search engines provided by Google, Microsoft and Yahoo! for the Chinese market along with the domestic Chinese search engine Baidu. This report finds that although Internet users in China are able to access more information due to the presence of foreign search engines the web sites that are censored are often the only sources of alternative information available for politically sensitive topics. In addition to censoring the web sites of Chinese dissidents and the Falun Gong movement, the web sites of major news organizations, such as the BBC, as well as international advocacy organizations, such as Human Rights Watch, are also censored. The data presented in this report indicates that there is not a comprehensive system - such as a list issued by the Chinese government - in place for determining censored content. In fact, the evidence suggests that search engine companies themselves are selecting the specific web sites to be censored raising the possibility of over blocking as well as indicating that there is significant flexibility in choosing how to implement China's censorship requirements. Finally, this report finds that search engine companies maintain an overall low level of transparency regarding their censorship practices and concludes that independent monitoring is required to evaluate their compliance with public pledges regarding commitments to transparency and human rights.

> Google.cn presents to users a clear notification whenever links have been removed from our search results in response to local laws and regulations in China.[2] - Google
>
> Where a government requests that we restrict search results, we will do so if required by applicable law and only in a way that impacts the results as narrowly as possible. If we are required to restrict search results, we will strive to achieve maximum transparency to the user.[3] - Yahoo!
>
> When local laws require the company to block access to certain content, Microsoft will ensure that users know why that content was blocked, by notifying them that access has been limited due to a government restriction.[4] - Microsoft

[1] Nart Villeneuve is a PHD student at the University of Toronto and a Senior Research Fellow at the Citizen Lab at the Munk Centre of International Studies. I am grateful for the comments and suggestions provided by Ron Deibert, Colin Maclay, Derek Bambauer, Rebecca MacKinnon and Sarah Boland. This project was made possible by the support of the Citizen Lab, the Berkman Center for Internet & Society and Social Sciences and Humanities Research Council. The opinions expressed in this report are solely that of the author. The data used for the report is available at: http://www.nartv.org/projects/search_monitor/

[2] Schrage, E. (2006). "Testimony of Google Inc." *Joint Hearing of the Subcommittee on Africa, Global Human Rights & International Operations and the Subcommittee on Asia and the Pacific*. Retrieved, May 22 2008, from http://googleblog.blogspot.com/2006/02/testimony-internet-in-china.html

[3] Callahan, Michael. (2006). "Testimony of Michael Callahan." *Joint Hearing of the Subcommittee on Africa, Global Human Rights & International Operations and the Subcommittee on Asia and the Pacific*. Retrieved, May 22 2008, from http://yhoo.client.shareholder.com/releasedetail.cfm?releaseid=187725

Search engines are increasingly tailoring their results to exclude politically sensitive content, often by geographic location. This development has a significant, negative impact on the right to freedom of expression. The most advanced case of censorship targeting political content occurs in search engines that market a specific version of their product for Internet users in China. Google, Microsoft and Yahoo! all maintain versions of their search engines for the Chinese market that censor political content. In addition to the removal of content widely acknowledged as useful and credible, the censorship process lacks transparency and accountability. Testifying before the U.S. Congressional Subcommittee on Africa, Global Human Rights & International Operations and the Subcommittee on Asia and the Pacific in 2006, representatives from Google, Microsoft and Yahoo! all pledged to maintain or increase the levels of transparency and accountability with regard to their censorship practices.

Through empirical investigations into the actual practices of these companies, the Search Monitor Project compares the level of transparency and censored content across the search engines provided by Google, Microsoft and Yahoo! for the Chinese market. The analysis of these results is used to interrogate the significance of censored content and the process that determines what content is censored. The project aims to provide the basis upon which following questions may be addressed:

- How transparent are the censored search engines provided by Google, Microsoft and Yahoo!?
- How do they vary amongst themselves and how do they compare with domestic Chinese search engines? Does their implementation of filtering match public commitments they've made?
- How does the process of search engine censorship work? Does China order the search engines to block specific content? Do the search engines interpret general guidelines?
- Are Chinese citizens better off with the censored services of these search engines?

Summary

- *Transparency*: While Google, Microsoft and Yahoo! all provide some form of notification indicating that the versions of their search engines for the Chinese market are censored, each implements the notification in a different way. Despite public pressure and ongoing efforts to create a code of conduct for operating in censored environments, the overall level of transparency has actually declined in the cases of Microsoft and Yahoo! between 2006 and 2008. While Google has held steady in maintaining a higher degree of transparency, no further improvement has been made. The low level of transparency impedes the ability to closely monitor and compare the censorship practiced by these search engines.

---

[4] Krumholtz, J. (2006). "Congressional Testimony: The Internet in China: A Tool for Freedom or Suppression?" *Joint Hearing of the Subcommittee on Africa, Global Human Rights & International Operations and the Subcommittee on Asia and the Pacific*. Retrieved, May 22 2008, from http://www.microsoft.com/presspass/exec/krumholtz/02-15WrittenTestimony.mspx

- *Process*: Google, Microsoft, Yahoo! and the domestic Chinese search engine Baidu censor significantly different content. The low overall overlap among all four search engines indicates that there is not a comprehensive system (such as a list issued by the Chinese government) in place for determining censored content. In fact, the evidence suggests that the search engine companies themselves are selecting the specific web sites to be censored. The lack of consistency raises the possibility that these search engines may be engaged in anticipatory blocking which raises the possibility of over blocking.[5] This does not rule out the possibility that China may be providing guidance, in some of form, concerning content, or categories of content, to be censored. However, it also indicates that search engine companies have significant flexibility in choosing how to implement China's censorship requirements.[6] The lack of clarity in the process and the unwillingness of companies to disclose this information acts to bolster China's current censorship policy that thrives on secrecy and unaccountability.

- *Content*: Tests conducted between November 2007 and April 2008 show that 33%, or 130 of the 393, web sites returned from the search queries in each test run were censored by at least one search engine.[7] Google maintained the lowest average number of censored sites at a rate of 15.2% and was closely followed by Microsoft 15.7%. Baidu ranked the highest at 26.4% and Yahoo! averaged 20.8%. Consistently blocked content focused on news and dissident web sites, human rights groups, sites related to the Falun Gong movement, and pornography. There were significant fluctuations in censored content over time and each search engines censored different content. The results indicate that Internet users in China are able to retrieve a slightly wider array of content (20% more, on average)[8] due to the presence of foreign search engines.

- *Significance*: Although the total number of censored sites is not high, especially when compared to the amount of indexed sites, the significance of these sites in providing alternative information should not be underestimated. These censored sites are often the only sources of alternative information available in the top ten results for politically sensitive search queries. Moreover, even the uncensored versions of these search engines highly rank content that is hosted in China or ends in the domain suffix .cn, both of which China retains control over and are thus unlikely to present alternative information. Although, these search engines censor less content than the domestic Chinese search engine Baidu, the removal of these sites from the search engines has an unambiguous, negative impact on the freedom of expression.

---

[5] I am thankful to Derek Bambauer for raising this particular issue.
[6] I am thankful to Rebecca MacKinnon for raising this important point.
[7] Each web site returned from a query in an uncensored censored engine was tested in the censored versions of Google, Microsoft and Yahoo! as well as Baidu. For more information on methodology see Appendix A.
[8] When the results from Google, Microsoft and Yahoo are combined, 20% of the sites censored by Baidu are available. However, individually they provide more information, especially Google and Microsoft which provide, on average, 51% and 55% more content (content not available in Baidu) while Yahoo! averages 25% more.

- *Monitoring*: Independent monitoring is required to empirically establish levels search engine censorship and evaluate compliance with public pledges regarding commitments to transparency, accountability and human rights. This helps prevent backsliding on the part of search engine companies as well as ameliorate any misleading charges levied against them. It also allows companies to access information concerning their competitors' practices that would not otherwise be revealed. An accurate account of search engine censorship is a step toward demystifying and exposing China's Internet censorship policies.

Search engines have become the premier gatekeepers of the Internet. All over the globe, Internet users rely on a handful of search engines to find content that is most relevant to the key words used as queries. Beyond seeking to provide the most locally relevant results, these search engines are actively removing specific sites from their localized versions to comply with local laws around the world. While most of the focus is on hate speech, (child) pornography and copyright issues, search engines also act to censor political content. The most advanced case of such censorship concerns search engines that market a version of their product in China. Google, Microsoft and Yahoo! have all been severely criticized for their participation in the violation of the rights and freedoms of Chinese Internet users.[9]

Corporations are beginning to frequently face the "thorny ethical problem" of having to engage in behaviour that is "squarely at odds with the law, norms or ethics of the corporation's home state."[10] China, for example, has implemented a complex information security and censorship strategy that involves a web of legal restrictions and regulations combined with advanced technical content filtering/blocking and surveillance mechanisms.[11] This has created a climate of self-censorship that thrives on secrecy and unaccountability in which technology companies act to restrict their own content to comply with China's complex censorship policies.[12] In response to growing "bottom-up" criticism from share holders, writers, activists and Internet users both inside and outside China along with "top down" pressures from the U.S. Congress and the European Parliament, companies such as Google, Microsoft and Yahoo! have pledged to increase levels of transparency and minimize the impact on freedom of expression by narrowly interpreting China's censorship requests. Faced with the paradox of having to follow conflicting local laws, those of China requiring censorship and those of the U.S.

---

[9] Human Rights Watch. (2006). Race to the Bottom: Corporate Complicity in Chinese Internet Censorship. *Human Rights Watch*. Eds. R. MacKinnon et al. (18,8 (C)). Retrieved, May 22 2008, from http://www.hrw.org/reports/2006/china0806/

[10] Palfrey, J. and Zittrain, J. (2007). Catalysts for corporate responsibility in cyberspace. *Cnet News*, August 14, 2007. Retrieved, May 22 2008, from http://www.law.harvard.edu/news/2007/08/14_palfrey.php

[11] OpenNet Initiative. (2008). China (including Hong Kong). *Access Denied : The Practice and Policy of Global Internet Filtering*. Eds. R. Deibert, J. Palfrey, R. Rohozinski, J. Zittrain. Cambridge, MA: MIT Press. Retrieved, May 22 2008, from http://opennet.net/research/profiles/china

[12] Reporters Without Borders. (2002). Open letter to the Yahoo! chairman. Retrieved, May 22 2008, from http://www.rsf.org/article.php3?id_article=2959, see also http://opennet.net/studies/china/ and http://www.theatlantic.com/doc/200803/chinese-firewall

potentially requiring open access, search engine companies and other technology corporations are opting for a form of industry self-regulation.

A group of civil society organizations and major corporations formed, with the facilitation of the Business for Social Responsibility, to develop a code of conduct in an effort to guide the behaviour of corporations when faced with laws that interfere with human rights.[13] While the process is still ongoing it is not expected to be a "corporate pledge of civil disobedience" but will instead "focus primarily on transparency and accountability around privacy and censorship."[14] One of the key components in the process is to develop mechanisms to "hold signatories accountable."[15]

Without meaningful mechanisms to monitor and evaluate compliance there is always the risk that corporate social responsibility will be interpreted as mere public relations, particularly when codes of conduct emerge after episodes of intense criticism.[16] In order to be effective, external monitoring is required to ensure that corporations comply with their public pledges. As Jonathan Zittrain and John Palfrey argue:

> A critical part of such a voluntary process to establish a code, regardless of its substantive terms and who drafted it, is to develop an institution charged with monitoring (and ideally supporting through best practices) adherence to the code and pointing out shortcomings.[17]

The same code may be interpreted and implemented differently by each participating corporation making it difficult to determine the overall impact of such codes on improving human rights. Therefore, it is critical to engage in comparisons across corporations providing similar services.[18]

Independent monitoring that accurately interrogates search engine censorship and evaluates search engine companies' compliance with their public pledges is an integral component in preventing possible backsliding. It also acts to clarify the practices of these companies and can ameliorate misleading charges levied against search engine companies. An accurate account of search engine censorship is also a necessary step in demystifying and exposing China's Internet censorship policies.

---

[13] Palfrey, J. (2007). Reluctant Gatekeepers: Corporate Ethics on a Filtered Internet. *GLOBAL INFORMATION TECHNOLOGY REPORT*, p. 69, World Economic Forum, 2006-2007. Retrieved May 22 2008 from http://ssrn.com/abstract=978507

[14] Mackinnon, R. (2007). Shi Tao, Yahoo!, and the lessons for corporate social responsibility. Retrieved May 22 2008 from http://rconversation.blogs.com/YahooShiTaoLessons.pdf

[15] Baue, B. (2007). "From Competition to Cooperation: Companies Collaborate on Social and Environmental Issues". *Sustainability Investment News*. Retrieved May 22 2008, from http://www.socialfunds.com/news/article.cgi/2208.html

[16] Addo, Michael K. (1999). "Human Rights and Transnational Corporations - An Introduction." *Human rights standards and the responsibility of transnational corporations*. Kluwer, p. 11.

[17] Palfrey, J. and Zittrain, J. (2007). Catalysts for corporate responsibility in cyberspace. *Cnet News*, August 14, 2007. Retrieved, May 22 2008, from http://www.law.harvard.edu/news/2007/08/14_palfrey.php

[18] McLeay, Fiona. (2006). "Corporate Codes of Conduct." *Transnational Corporations and Human Rights*, Olivier De Schutter ed. Hart Publishing, p. 231.

The Search Monitor Project uncovers and compares the censorship practices of the search engine services that Google, Microsoft and Yahoo! operate for the Chinese market. The first component of the project focuses on transparency, in particular, on the presence or absence of notification indicating censorship. The second examines the censorship process by comparing the frequency of censored web sites in relation to key words that are used as queries across each of the search engines along with the domestic Chinese search engine, Baidu. It also compiles and compares censored web sites across all the engines. The third component analyzes censored content by examining content that is censored across all search engines. It also provides a comparison between the Chinese-language "global" versions of Google and Yahoo and their censored China-specific versions. Organized in this way the results raise questions regarding the nature of censorship process as well as the censored content.

Transparency

In 2006, Google, Microsoft and Yahoo! introduced a message that informed users when the results of their searches were censored. The presence of a mechanism of notification is a critical component of transparency. This notification informs users that their search results have been censored and indicates, to a certain degree, the reason (often unspecified "local law") why the results have been censored.

While all three companies publicly committed to such notification they differ considerably in terms of implementation. In addition, between 2006 and 2008 the level of transparency, overall, has actually decreased.[19] While Google's censorship notification has remained the same as it was in 2006, Yahoo! and Microsoft have altered the way in which users are notified of censorship. Yahoo! has put its censorship message at the bottom of every page regardless of whether results are censored or not, in effect de-linking the censorship notification from the results. Microsoft removed the text from the results page completely and buried the censorship notification in a separate "help" page. However, Microsoft did restore the censorship notification to instances of particular search queries, but the notification was not restored when searches are restricted to a particular censored website. These developments represent a significant degrading of transparency and accountability.

The Search Monitor Project assesses these notifications based on four components:

- Presence: The presence of a mechanism of notification that informs users that their results may be censored.
- Placement: The location of the censorship notification message, particularly its placement in relation to the results.

[19] This project focuses on the notification that appears when web sites are de-listed from search results. There have been some recent changes to the search engines' notification concerning specific "key word" queries. For example, certain queries are restricted and return no results, just a censorship notification. These developments suggest that further research is required focusing on specific queries as well as de-listed web sites.

- Specificity: The extent to which users are informed about specific laws, orders and/or regulations leading to censored results.
- Connection: Notification appears only when content is actually removed in relation to what the user searches for making it possible to determine which specific web sites and keywords have actually been censored.

The failure to include any form of censorship notification, or hiding the placement of the censorship message, creates a condition in which users may be unaware that their results have been censored. Furthermore, by de-linking the censorship notification from the queries and/or results (by for example, displaying the censorship notification regardless of what the user actually searched for), the topics and websites that are censored remain hidden from the user. The de-linking of the censorship message from the search results impacts the ability to determine what precise sites and "key words" are being censored.

The presence and placement of a censorship notification, along with the specificity of its content and its connection to the results, is an integral component of transparency. The specificity of the reason why content has been removed is an important component that is lacking in the case of China. In other cases, Google has cited specific laws, such as the DMCA, and other legal documents with which they must comply and reported the information, to some degree, to Chilling Effects.org.[20] Yahoo! China maintains a list of sites it censors for copyright violations.[21] However, in the case of censored political content in China nothing other than a reference to "local law" is provided.[22]

The presence of a notification that is directly connected to the results[23] positively impacts the ability to accurately identify censored website and restricted keywords. When such notifications are either absent or disconnected from the results (for example, a notification that appears on every page regardless of whether results are censored or not) the ability to determine censored sites with a high degree of confidence diminishes as sites may simply not be indexed by the search engine. Therefore, the notification is critical not only for informing users but also for the monitoring process.

June 26, 2006				
Engine	Presence	Placement	Specificity	Connection
Google	Yes	High Notification is placed under results	Low Results removed to comply with local law	Yes Notification only appears when results are censored

[20] See http://www.google.com/dmca.html
[21] See http://search.help.cn.yahoo.com/h3_9.html
[22] Human Rights Watch. (2006). Race to the Bottom: Corporate Complicity in Chinese Internet Censorship. *Human Rights Watch*. Eds. R. MacKinnon et al. (18,8 (C)). Retrieved, May 22 2008, from http://www.hrw.org/reports/2006/china0806/5.htm#_Toc142395824
[23] This refers to notification that appears only when content is removed in relation to what queries the user enters into the search engine.

Yahoo!	Yes	High* Notification is placed under results	Low Results removed to comply with local law	Yes*
Microsoft	Yes	High Notification is placed under results	Low Results removed to comply with local law	Yes Notification only appears when results are censored

May 13, 2008				
Engine	Presence	Placement	Specificity	Connection
Google	Yes	High Notification is placed under results	Low Results removed to comply with local law	Yes Notification only appears when results are censored
Yahoo!	Yes	Medium Notification is placed at the bottom of every page	Low Results removed to comply with local law	No
Microsoft	Yes**	Medium Notification when searching for particular "key words".**	Low Results removed to comply with local law	Yes**

* Yahoo China's web crawlers operate from within China, behind the China's filtering system, therefore sites that are blocked by China are not indexed by Yahoo (and thus do not need to be censored by Yahoo) leaving only sites that are either not blocked by China or are indexed during periods when there is variation in the capacity of China's filtering system to actually be censored by Yahoo. The behaviour documented here refers to sites indexed by Yahoo but subsequently censored, not sites that are not indexed by Yahoo at all.

** Microsoft provides notification when searching for particular "key words", however, no message appears when restricting the search to a censored web site. It is therefore difficult to determine with precision that a specific website has in fact been censored.

While Google, Microsoft and Yahoo! all provide some form of notification indicating that the versions of their search engines for the Chinese market are censored, each implements the notification in a different way. Despite public pressure and ongoing efforts to create a code of conduct the overall level of transparency has actually declined in the cases of Microsoft and Yahoo!. While Google has held steady in maintaining a higher degree of transparency, no further improvement has been made.

Methodology

Building upon previous research conducted by Reporters Without Borders and Human Rights Watch, the Search Monitor Project compares the level of censorship across the search engine services that Google, Microsoft and Yahoo! censor for the Chinese market

as well as the domestic Chinese search engine, Baidu.[24] A set of sixty keywords have been selected covering the broad topical categories of censorship circumvention, the Falun Gong movement, political sensitivities and social taboos. The keywords (the majority of which are in Chinese) have been selected in order to uncover censored sites they consist of specific topics as well as general words and phrases.

These keywords are used as search queries in uncensored search engines. The web sites (URLs) returned from the uncensored search engines is used to build a list of unique web sites (domain names). These web sites are checked in the censored search engines using the "site:" modifier in order to restrict the results set to pages from the specific web site being tested.

In cases where the censored search engine being tested displays a censorship notification message that only appears when results have been censored, domains that produce no results when queried with the "site:" modifier and contain a "censor message" are labeled as "Censored." In cases where domains return some results but also contain a "censor message" they are labeled as "Page Censored" indicating that partial content is available. In the cases where there is no censor message, or the censorship message appears on every page and bears no connection to the results, domains that produce no results when queried with the "site:" modifier are labeled as "Censored."[25]

Google and Microsoft host versions of their search engines outside of China[26] and operate their "web crawlers", the software that indexes Internet content, from outside of China. However, Yahoo! and Baidu host their search engines inside China and operate their "web crawlers" from inside China. This is significant because China's filtering system, often called the Great Firewall of China (GFW), can interfere with and block requests to search engines that contain particular keywords. Since the goal of this project is to test the levels of censorship of the search engines themselves it is necessary to "ignore" the filtering conducted by China. As Clayton, Murdoch and Watson reveal, Internet traffic to and from China passes through a filtering system that is bi-directional - it affects both inbound and outbound traffic - which disrupts connections if the presence of particular key words are detected.[27] Often, China will designate a domain name as

---

[24] This methodology not only builds upon the successful aspects of previous research conducted by RSF and HRW but can hopefully explain some of the anomalies previously identified and avoid potential pitfalls. See Appendix A for a full description of the methodology. For RSF's report, see http://www.rsf.org/article.php3?id_article=18015. For HRW's report, see http://www.hrw.org/reports/2006/china0806/.

[25] As explained in Appendix A, it is important to note that sites that are simply not indexed by the search engine will appear as "Censored" thus possibly inflating the total amount to censorship attributed to search engines that do not have a censor message that is related to the results. This can be slightly compensated for by looking at the overlap of censored sites among search engines. In addition, since this is a normative project advocating transparency, should serve as an incentive for search engines to implement a censor message that is related to the results.

[26] Some servers for google.cn are hosted inside China, but some google.cn servers are located outside China and can be queried from outside China.

[27] Clayton, R., Murdoch, S. and Watson R. (2006, June 28 - June 30). Ignoring the Great Firewall of China. *Paper presented at the 6th Workshop on Privacy Enhancing Technologies*, Cambridge, United Kingdom. Retrieved May 22 2008, from http://www.cl.cam.ac.uk/~rnc1/ignoring.pdf

"key word" thus disrupting access for any request that contains that domain name. This is important as queries directed to search engines hosted in China use the "site:" modifier followed by a domain name.

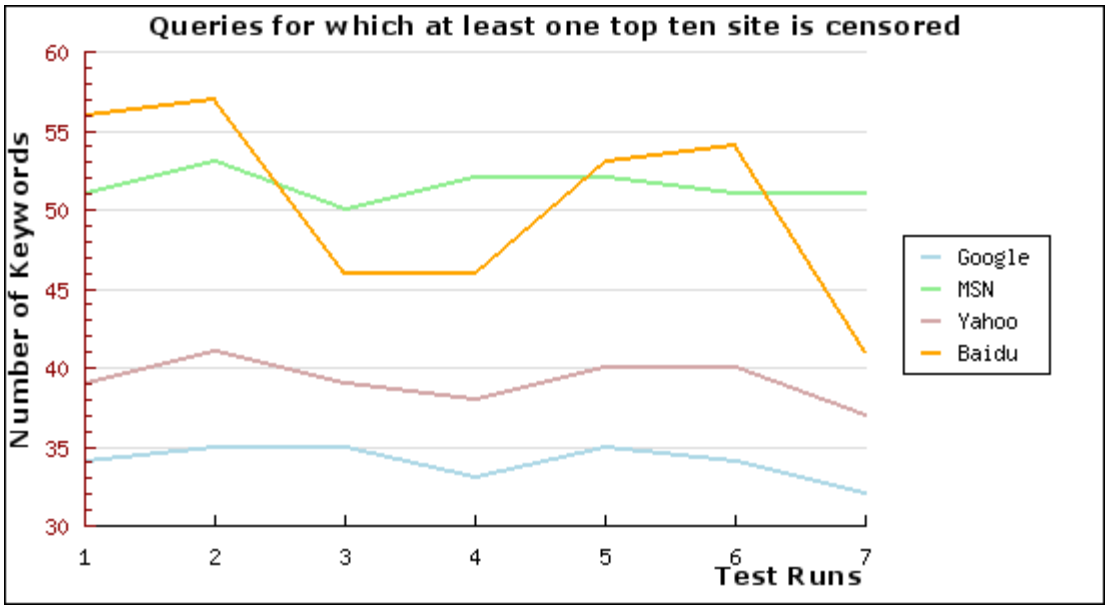
Another important factor that results form China's filtering system concerns the search engines that operate their crawlers from behind the GFW. Yahoo! and Baidu both operate their search spiders from inside China. This results in a situation where, because of China's gateway filtering, the crawlers that index content for these search engines cannot access sites that China blocks. Thus they rarely have to de-list specific websites, since most are just not indexed in the first place.[28] In order to avoid interference from the China's filtering system, the China-specific versions of Google[29] and Microsoft, which are hosted outside of China, are queried from outside of China. The China-specific versions of Yahoo and Baidu, which are hosted inside China, are queried from inside China. In this way, the requests to the search engines do not pass through the GFW.

This methodology allows the amount of keywords for which there are censored results as well as the total amount of censored web sites to be compared across search engines and tracked over time.

Results

The following results are based on data compiled from seven test runs that took place between November 2007 and April 2008.[30] On average, 43 of the 60 keywords (71%) used to generate lists of web sites resulted in at least one censored site in the top ten returned results.[31]

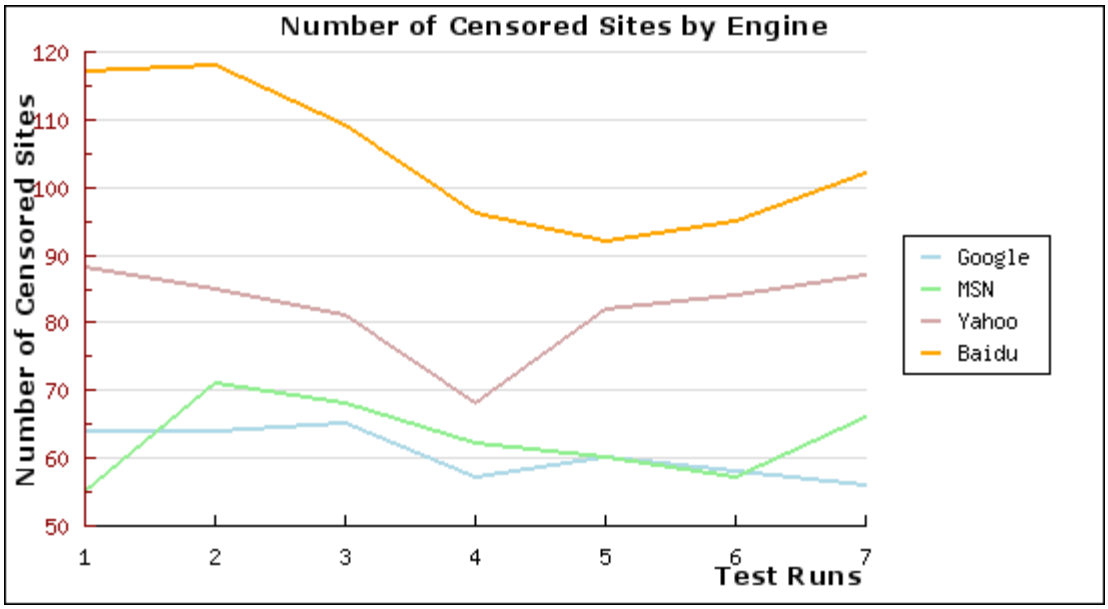
[28] However, this also leads to situations in which sites that are normally blocked by China, and de-listed by Google and Microsoft, are indexed by Yahoo!. The GFW is not 100% effective and occasionally crawlers operating from inside China are able to index a normally blocked site which then appears in their search results. It is also important to note that sites indexed by the search engines that are blocked by the GFW will still be inaccessible to users in China. Also, keywords used to query Google and Microsoft from outside China, may be blocked by the GFW preventing users from inside of China from receiving results.
[29] Some servers for google.cn are hosted inside China, but some google.cn servers are located outside China and can be queried from outside China.
[30] One test run was completed using the "uncensored" Yahoo! search engine to generate the URL result set, however, it has been excluded from the data analysis presented here. All the test runs are based on queries made to the "uncensored" google.com.
[31] It is important to note that these keywords are not censored by the search engines. These statistics refer to keywords used in "uncensored" search engines to generate a set of web sites (URLs) that are tested in the censored versions of these search engines. If at least one of these web sites is found to be censored, the keyword count is incremented by one. With the exception of Google, there is no reliable way to determine if specific keywords are being censored. This report is focused on search engine comparison; therefore, testing for censored keywords has been excluded. Future studies will focus on key word censorship itself.



Google maintained the lowest average with 34/60 (56.6%) while Microsoft returned the highest with an average of 51/60 (85%). Yahoo! maintained an average of 39/60 (65%) and Baidu 50/60 (83.3%). This highest count of keywords on a single test run was recorded by Baidu with 57/60 (95%) and Google returned the lowest with 32/60 (53.3%).

These results indicate that it is not uncommon for results to be censored when searching for political content. While specific queries such as “Epoch Times” (大紀元) produce consistently censored web sites, the results indicate that searches for generic queries such as “human rights” (人权) or “democracy” (民主) yield censored web sites as well.

However, there has been an overall decline in the number of keywords that produce at least one censored result. While this may indicate an increased focus on particular content areas and a decline in others, it most likely represents the fluctuations that occur with search engine rankings.



The following displays the results of each web site checked in the censored version of the search engines. On average 130 of the 393 web sites (33%) returned from the search queries were censored by at least one search engine. Google maintained the lowest average number of censored sites per test run with 60/393 (15.2%) and was followed by Microsoft at 62/393 (15.7%). Baidu ranked the highest at 104/393 (26.4%) and Yahoo! averaged 82/393 (20.8%). The highest single count of censored web sites was recorded by Baidu at 118/401 (29.4%) while Microsoft recorded the lowest with 55/387 (14.2%).[32]

The results indicate that there are fluctuations in the levels of censorship. However, it is unclear whether this is a result of actual changes in censorship practices or fluctuations in the ranking of results. However, the results do indicate that search engines hosted outside (Google, Microsoft) of China censor considerably less than those hosted inside China (Yahoo!, Baidu).

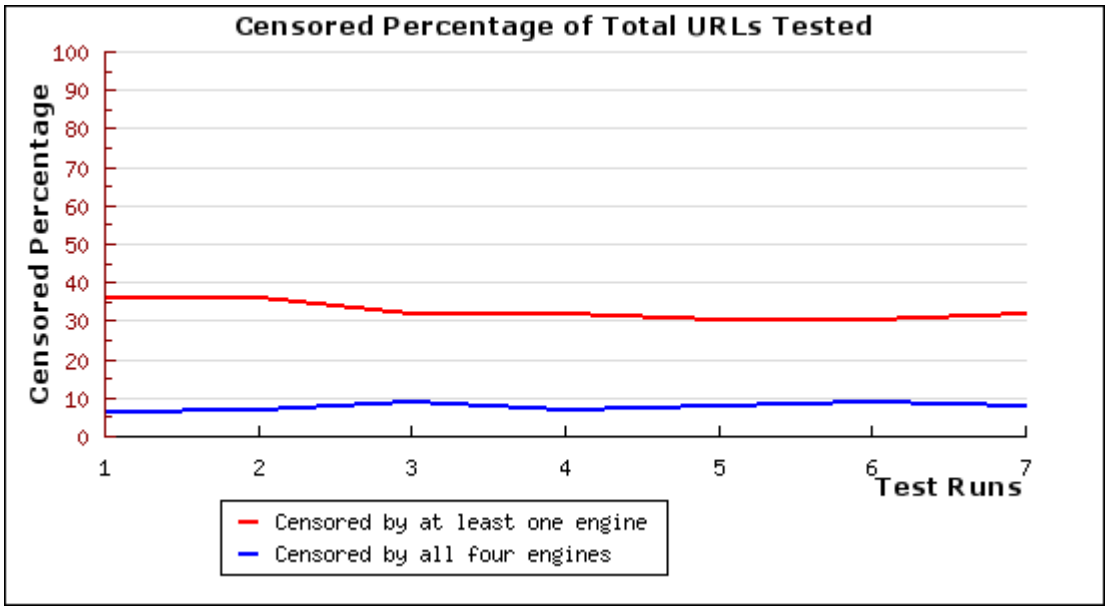
Process

The way(s) in which censored content is determined remains unclear. The role of the Chinese government in the process of search engine censorship is uncertain. It is not known if search engine companies are explicitly given notice of the exact content to block or if these companies simply infer what the Chinese government would like to have blocked and determine that content themselves. A New York Times report indicates that "self discipline" is the major driving force of censorship in China. The article contends that it is companies themselves - not the government of China - that are deciding what specific content to block.

[32] This result was obtained before it was discovered that Microsoft does not properly handle links that begin with "https." Sites beginning with "http" are properly censored, but any links to them that begin with "https" – remain. Therefore results that only contain links beginning with "https" are considered partially censored, or "PageCensored". After adjustment for this development Google ranks the lowest with 56/392 (14.2%).

> American Internet firms typically arrive in China expecting the government to hand them an official blacklist of sites and words they must censor. They quickly discover that no master list exists. Instead, the government simply insists the firms interpret the vague regulations themselves.[33]

In an effort to provide some insight regarding the question of process, the Search Monitor Project analyzes the overlap of subsets of censored content among search engines that are functionally similar (Google/Microsoft, Yahoo/Baidu). Overlap refers to the sites that are censored by multiple search engines. Overlap is analyzed in three ways. The first focuses on sites that are censored by all four search engines tested. The second focuses on search engines that censor using similar mechanisms. In this case, Google and Microsoft are paired together and Yahoo! and Baidu are grouped together. The third focuses on the overlap between the two groups of search engines. While this allows for a comparison among search engines it also acts as an indicator of whether the search engines are responding to specific blocking requests, usually associated with an official order, or a general determination on the part of company, perhaps based on topic areas provided by officials.

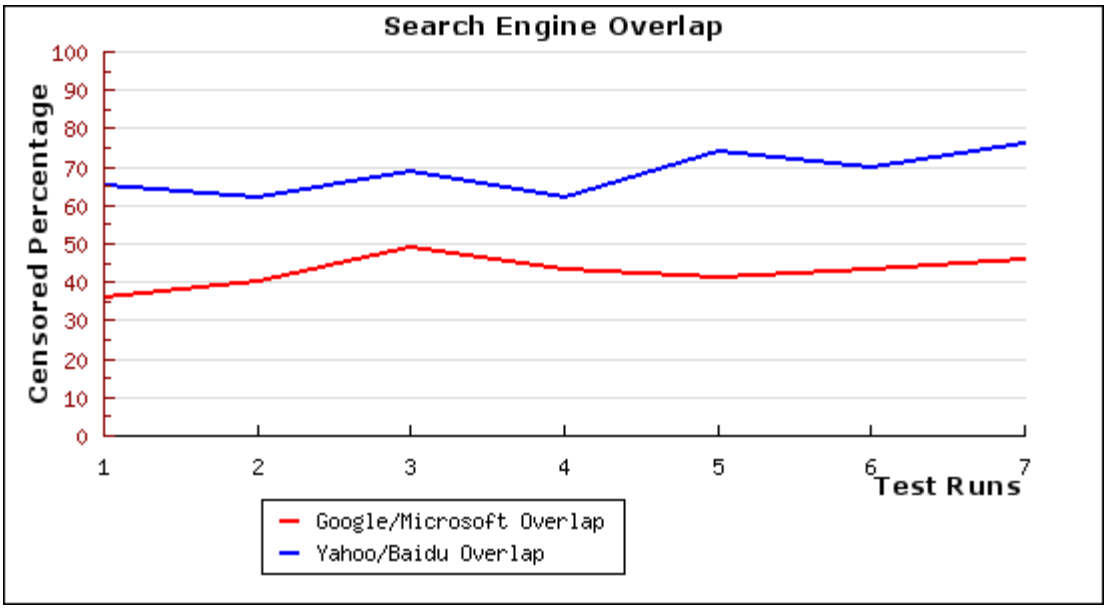


These statistics represent the number of sites censored by all four search engines divided by the total number of sites tested and the number of sites censored by at least one search engine divided by the total number of sites tested. On average, 8.1% of the total number of web sites tested was censored by all four search engines. However, on average, 24.6% of the total number of web sites tested was censored by at least one search engine. The highest instance of overlap among all four search engines was 9.4 while the lowest was 6.7%. The highest instance of websites censored by at least one search engine was 29.1 of

[33] Thompson, C. (2006). Google's China Problem (and China's Google Problem). *The New York Times*, April 23, 2006. Retrieved May 22 2008, from http://www.nytimes.com/2006/04/23/magazine/23google.html

the total number of sites tested while the lowest was 18.1%. While remaining relatively stable over time, the amount of sites censored by all four search engines remains low. This indicates that they are censoring significantly different content.

While the low overall overlap among all four search engines appears to indicate that there is not a comprehensive system for determining censored content, it also may simply reflect the differences in implementation among the search engines. For example, search engines could choose to block some but not all content based on a centralized list. The lack of transparency among most search engines (no direct link between censorship notification and returned results) results in hinders the effort to establish a precise level of overall overlap because sites that are simply not indexed (no results) may be counted as censored sites since there is no reliable way to distinguish one from the other.[34] However, by grouping the overlap scores by functionally similar search engines this disparity should be corrected.

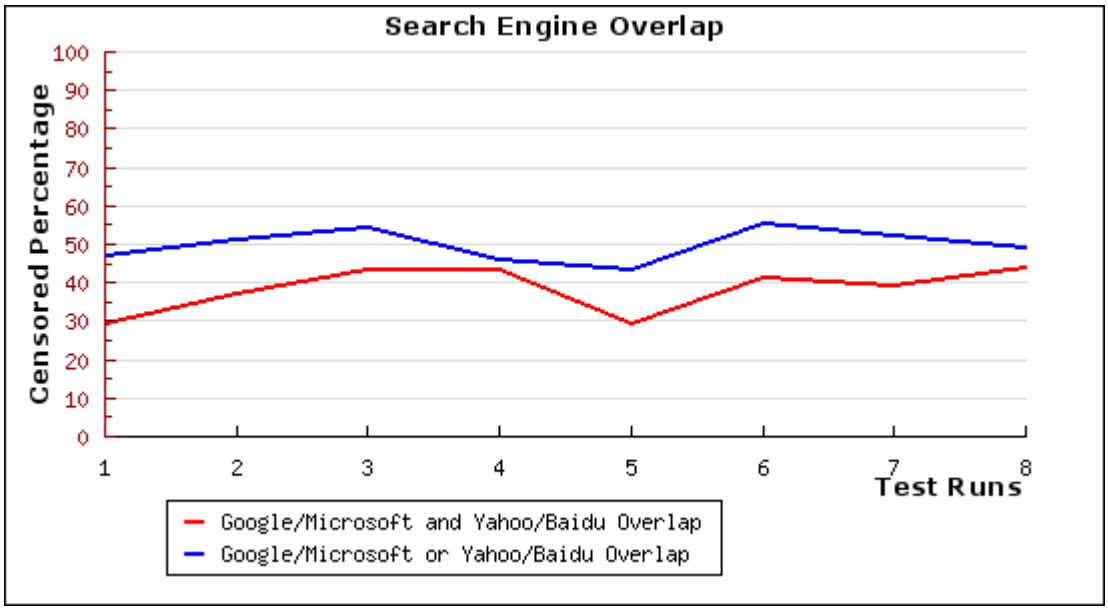


These statistics represent the count of web sites censored by both search engines, divided by the count of websites censored by either engine. For example, the count of websites censored by Google **and** Microsoft divided by the count of websites censored by Google **or** Microsoft. On average, the overlap of the websites censored by Google and Microsoft occurred at a rate of 37/86 (43%) while Yahoo! and Baidu recorded a rate of 75/110 (68.1%). The highest occurrence of overlap between Google and Microsoft occurred at a rate of 44/89 (49.4%) and the lowest was recorded at 32/87 (36.7%). The highest overlap among Yahoo! and Baidu was 82/107 (76.6%) and the lowest was 63/101 (62.3%).

Grouping the results by functionally similar search engines dramatically increases the overlap scores. However, they still indicate that the search engines are censoring different content. The evidence suggests that both Google and Microsoft are determining the

[34] See Appendix A for detailed description of methodology and a discussion of this issue.

precise (domain name) sites to be censored – or, alternatively, what precise sites not to censor. The overlap between Yahoo! and Baidu is significantly higher due to the fact that both operate their crawlers from behind China's filtering system and so both should not be able to index sites that China blocks. Although the remaining disparity can be partially explained by Baidu's failure to index as heavily as Yahoo!, it suggests that some additional censorship is likely being conducted by the search engine companies themselves.[35]



In order to further explore the range of censored content, the combined results from the pairs of functionally similar search engines are analyzed. The first data set reflects the overlap rate of sites that both Google and Microsoft censor with the sites that both Yahoo! and Baidu censor. The second shows the overlap rate of sites censored by Google or Microsoft and Yahoo! or Baidu.

On average the overlap rate of Google/Microsoft and Yahoo/Baidu stands at 32/83 (38%) while the average for Google/Microsoft or Yahoo/Baidu was 66/132 (50%). The highest level of overlap for the Google/Microsoft and Yahoo/Baidu set was 37/84 (44%) and the lowest was 31/106 (29%). The highest level of overlap for the Google/Microsoft or Yahoo/Baidu set was 71/130 (54%) and the lowest was 66/151 (43%).

The results indicate that the search engines hosted outside of China (Google and Microsoft), which implement censorship by de-listing results, censor significantly different content than those hosted inside China (Yahoo! and Baidu), which do not index sites blocked by China's filtering system.

[35] There are web sites which are indexed by Yahoo which are not censored by China's filtering system that are not indexed by Baidu.

Google, Microsoft, Yahoo! and the domestic Chinese search engine Baidu are censoring significantly different content. The low overall overlap among all four search engines indicates that there is not a comprehensive system (such as a list issued by the Chinese government) in place for determining censored content. In fact, the evidence suggests that the search engines companies themselves are selecting the specific web sites to be censored. The lack of consistency raises the possibility that these search engines may be engaged in anticipatory blocking which raises the possibility of over blocking.[36] This does not rule out the possibility that China may be providing guidance, in some of form, concerning content, or categories of content, to be censored. However, it also indicates that search engine companies have significant flexibility in choosing how to implement China's censorship requirements.[37]

Content

The following results are based on data compiled from seven test runs that took place between November 2007 and April 2008.[38] In total, 313 web sites (domains) were censored by at least one search engine on at least one occasion. However, there were only 76 web sites (domains) that were censored by all four search engines on at least one occasion. Due to the variations in search engine rankings, not all of these 76 sites were tested during each test run. The following analysis will focus on the 19 web sites (of the 76) that were tested on each occasion. Only 8 websites were found to be censored by all four search engines on each test run. There was variation among the levels of censorship of remaining 11 web sites across the test runs.

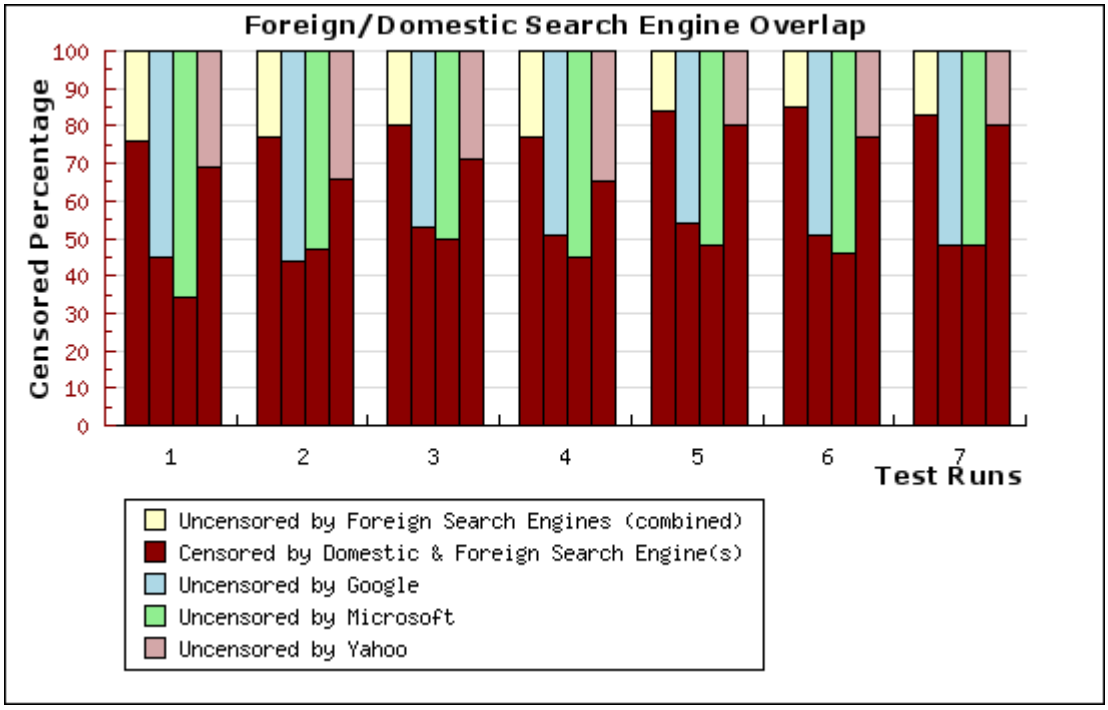
Domain	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7
www.falundafa.ca	C	C	C	C	C	C	C
www.falundafa.org.my	C	C	C	C	C	C	C
www.falundafa.org.tw	C	C	C	C	C	C	C
www.falundafa.org	C	C	C	C	C	C	C
www.fireofliberty.org	C	C	C	C	C	C	C
www.epochtimes.com	C	C	C	C	C	C	C
www.boxun.com	C	C	C	C	C	C	C
package.minghui.org	C	C	C	C	C	C	C
www.minghui.org	P	C	C	C	C	C	C
tw.fgmtv.org	P	C	C	C	C	C	C
tw.epochtimes.com	P	P	C	C	C	C	C
www.peacehall.com	C	P	P	P	C	C	C
www.voanews.com	C	C	C	P	P	P	P
web.wenxuecity.com	P	P	P	P	C	C	C
www.rfa.org	P	P	P	P	C	C	C
news.bbc.co.uk	P	P	P	P	C	P	P
big5.minghui.org	P	P	P	P	P	P	C
www.readxxx.com	P	P	P	P	P	P	C

[36] I am thankful to Derek Bambauer for raising this particular issue.
[37] I am thankful to Rebecca MacKinnon for raising this important point.
[38] One test run was completed using the "uncensored" Yahoo! search engine to generate the URL result set, however, it has been excluded from the data analysis presented here. All the test runs are based on queries made to the "uncensored" google.com.

minghui.org	P	P	P	P	P	P	C

* C = Censored by all four search engines
* P = Censored by at least one search engine

These sites can be roughly categorized as Falun Gong (9), news and dissidents (9), and pornography (1). Some other significant sites such as www.hrw.org and www.asiademo.org were censored by all four search engine on 6 test runs, but were not tested on one of the 7 test runs.



However, most content was not censored by all four search engines. Thus the diversity of available search engines increases the overall amount of information available to Chinese Internet users. The results indicate that Internet users in China are able to retrieve a slightly wider array of content, 20% more, on average, due to the presence of foreign search engines. When the results from Google, Microsoft and Yahoo are combined, 20% of the sites censored by Baidu are available. However, individually they provide more information, especially Google and Microsoft which provide, on average, 51% and 55% more content (content not available in Baidu) while Yahoo! averages 25% more.

Runs	1				2				3				4				5				6				7			
Domain	G	M	Y	B	G	M	Y	B	G	M	Y	B	G	M	Y	B	G	M	Y	B	G	M	Y	B	G	M	Y	B
www.rfa.org	C	C	I	C	C	C	I	C	C	C	I	C	C	C	I	C	C	C	C	C	C	C	C	C	C	C	C	C
news.bbc.co.uk	C	C	I	C	C	C	I	C	C	C	I	C	C	C	I	C	C	C	C	C	C	C	I	C	C	C	I	C
www.voanews.com	C	C	C	C	C	C	C	C	C	C	C	C	I	C	C	C	I	C	C	C	I	C	C	C	I	C	C	C
freenet-china.org	C	I	C	C	C	I	C	C	C	I	C	C	C	I	C	C	C	I	C	C	C	I	C	C	C	I	C	C
www.anonymizer.com	I	I	C	C	I	I	C	C	I	I	C	C	I	I	C	C	I	I	I	C	I	I	I	C	I	I	C	C
zh.wikipedia.org	I	C	I	C	I	C	I	C	I	C	I	I	I	C	I	I	I	C	I	C	I	C	I	C	I	C	I	I

* G = Google, M = Microsoft, Y = Yahoo, B = Baidu
* C = Censored, I = Indexed

This set of web sites was selected in order to demonstrate the variations in the availability of search engine results. Content that is generally unavailable to Internet users in China can be indexed by search engines.[39] These results indicate that Internet users in China are able to retrieve content censored by one engine through searches in another.

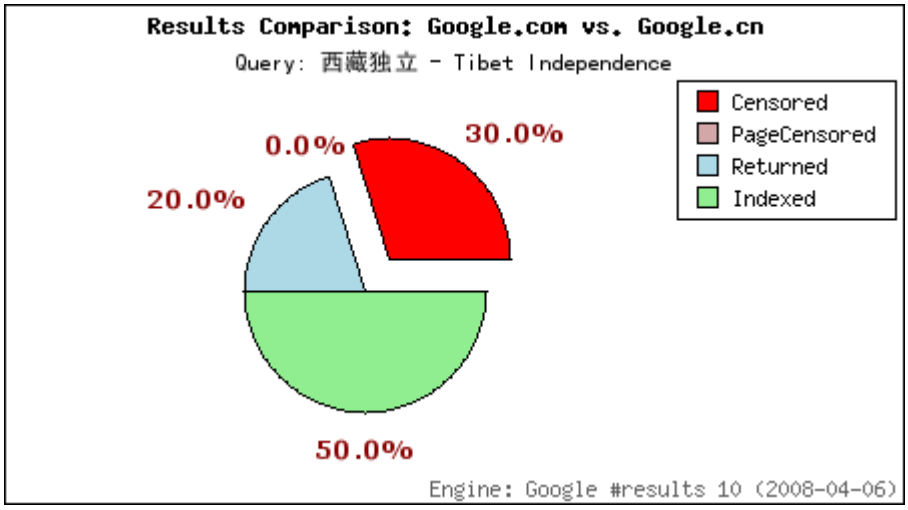
Significance

This component focuses on two of the search engines that have comparable censored and uncensored versions: Google and Yahoo!. It is a direct comparison between google.com (in Chinese) and google.cn and yahoo.com (in Chinese) and yahoo.cn. Rather than simply looking for censored content, this component also analyzes the overlap rate – the number of websites that appear in the top ten results in the censored and uncensored versions of the search engines. It also tracks which sites are hosted in China or end in a .cn domain suffix.

Since the total number of censored sites is relatively small compared to the total number of indexed sites, this component measures the significance of the censored sites in order to show just how important the censored sites are in relation to those displayed to the user. Significance refers to the number of top ten sites returned from an uncensored search engine that are censored in the China-specific version in relation to those that are either hosted in China or that end in a .cn domain suffix. China could, presumably, take action against those sites under their jurisdiction without having to resort to blocking. In this context, these sites are considered to be "authorized" and are unlikely to contain information that presents an alternative perspective to that approved by the government.[40]

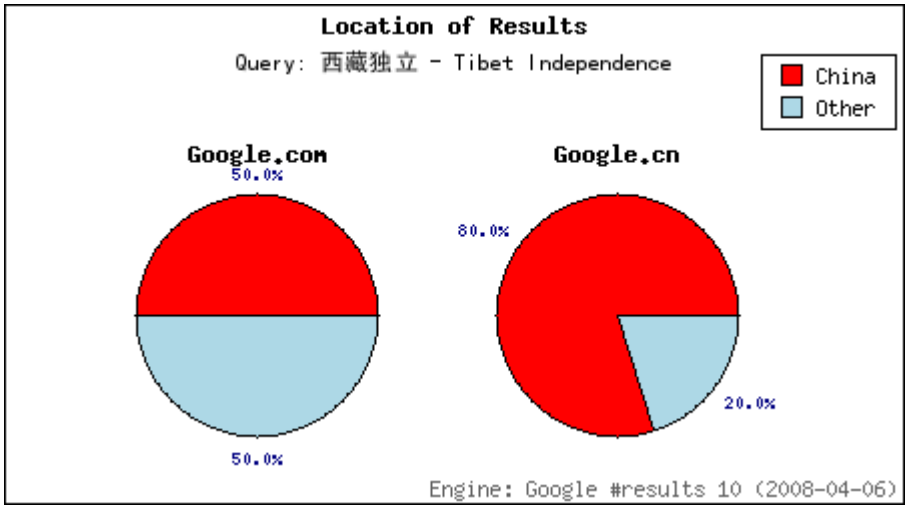
[39] China's filtering system is not 100% effective, and the software used by search engines occasion indexes sites that are usually blocked. These results may also reflect periods of blocking and unblocking.
[40] This is a rough measure as there may be sites that end in .cn (or even hosted in China) that may contain "unauthorized" information. However, China could apply controls that are easier and more effective than blocking to shut down such sites. In this component results that are returned in the top ten along with those that are indexed but not displayed in the top ten are distinguished from those that are censored.

The results presented here represent selected queries from the most recent test run (2008-04-06) for both Google and Yahoo!.



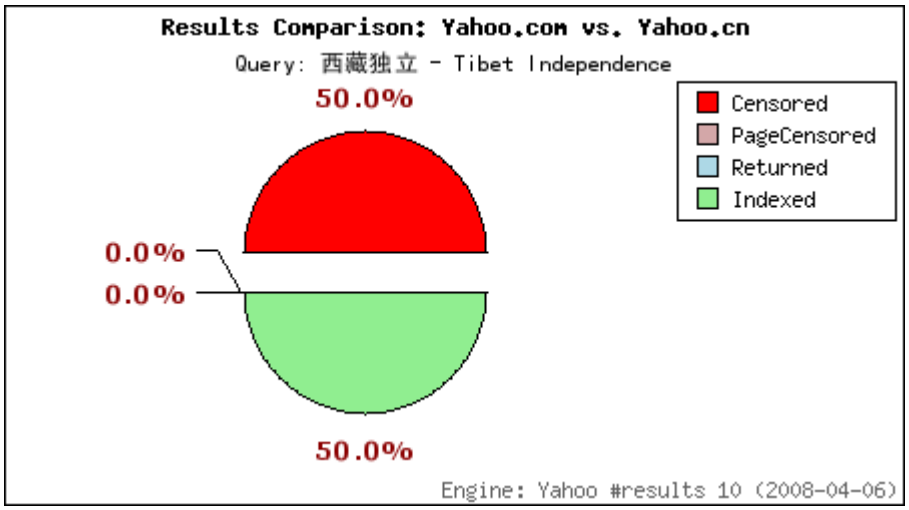
Only 2 of the top ten sites for the search query are the same in both google.com and google.cn. 3 are censored and 5 remain uncensored but do not appear in the top ten.

The censored sites are:
- www.tangben.com
- www.epochtimes.com
- www.tibetalk.com

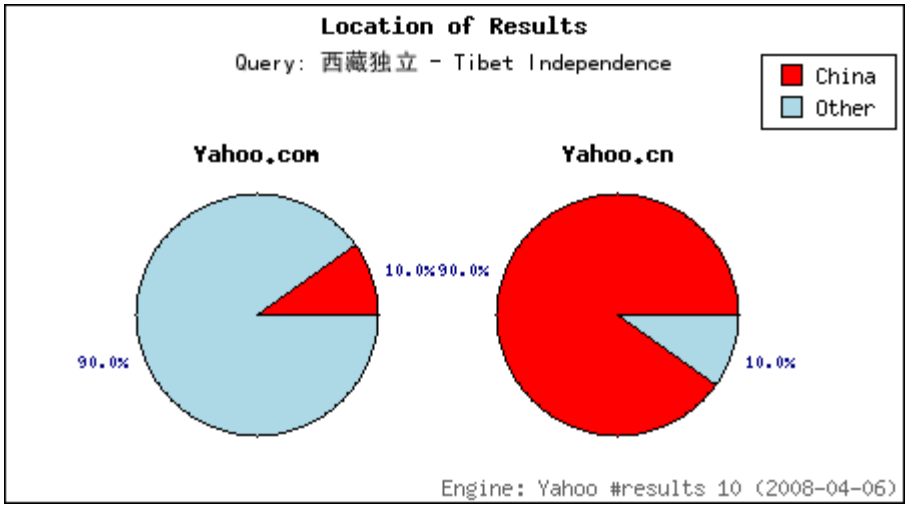


Despite only having 3 censored sites, 8 of the results in google.cn are "authorized" sites (they end in .cn or are hosted in China). Even in the uncensored google.com half of the top ten results represent "authorized" content.

These results indicate that although Google censors considerably less that the other search engines, "authorized" content is ranked high, even in the uncensored google.com. By prioritizing local content (as Google does withal other markets such as Canada's google.ca) the significance of the few censored sites is amplified as these are the only ones which represent divergent view points.



None of the top ten sites for the search query are the same in both yahoo.com and yahoo.cn. 5 are censored and 5 remain uncensored but do not appear in the top ten. The censored sites are:
- www.voanews.com (2)
- www.hkfront.org
- www.chinaaid.org
- www.rangzen.org



yahoo.cn displays significantly different results from those of yahoo.com. Fully 9 of the results in yahoo.cn are “authorized” sites (they end in .cn or are hosted in China). Notably, only one of results in the uncensored yahoo.com was “authorized” content.

It is quite clear that yahoo.com and yahoo.cn contain considerably different content. No top ten results returned by queries in yahoo.com were returned in the top ten results form the same query in yahoo.cn. While yahoo.com tends to not rank “authorized” content as highly as google.com, the results from yahoo.cn heavily favour “authorized” content.

Although the total number of censored sites may not be high, especially when compared to the amount of indexed sites, the significance of these sites in providing alternative information should not be underestimated. These censored sites are often the only sources of alternative information available in the top ten results for politically sensitive search queries.

Conclusion

It is becoming increasingly clear that technology companies face a dilemma when attempting to penetrate the Chinese market. A failure to comply with China's censorship policies can result in the wholesale blocking of a company's entire service or significant levels of interference due to China's filtering system. Companies that have a physical presence in China face the challenge of obtaining proper licensing and their Chinese employees may face legal threats for the foreign company's failure to comply with China's censorship policies. However, it is also clear that compliance with China's censorship policies is also an unattractive option. Google, Microsoft and Yahoo! are all facing tough criticism from governments, human rights groups and civil liberties advocates as well as their shareholders for their complicity in China's censorship policies.[41]

The empirical results presented in this study suggest that while the total amount of web sites censored by the search engines provided by Google, Microsoft and Yahoo! for the Chinese market may be relatively low, the significance of these sites is high. These censored sites are often the only sources of alternative information available in the top ten results for politically sensitive search queries. The removal of these sites from the results has an unambiguous, negative impact on the freedom of expression.

Since Google, Microsoft and Yahoo! have already made the choice to acquiesce to China's censorship policies the question of process becomes crucially important. The process of determining web sites to censor is important for establishing legitimacy. A transparent process minimizes potential abuse, enables mechanisms of oversight and allows for grievances to be filed. A process that is clouded in secrecy and wholly unaccountable is unacceptable and runs counter to the values of freedom and democracy that these companies profess.

While foreign search engines do provide more content than domestic search engines, the greatest benefit of having foreign search engines in China may not be increased access to information but is the potential contribution that these companies can make to further transparency and accountability in the process of censorship. China's current censorship policy is bolstered by secrecy and thrives on unaccountability. This is exemplified by the evidence which suggests that it is the search engine companies themselves, and not the government of China, that determine what content is to be censored. A transparent system would require the Chinese government to legally justify why content is censored and, as a consequence, provide for a mechanism of appeal.

However, overall level of transparency has actually declined in the cases of Microsoft and Yahoo! between 2006 and 2008. While Google has held steady in maintaining a higher degree of transparency, no further improvement has been made. The low level of

[41] For a detailed discussion of the dilemma companies' face and a comparison of the approaches Google, Microsoft and Yahoo! have taken, see Mackinnon, R. (2007). Shi Tao, Yahoo!, and the lessons for corporate social responsibility. Retrieved May 22 2008 from http://rconversation.blogs.com/YahooShiTaoLessons.pdf

transparency impedes the ability to closely monitor and compare the censorship practiced by these search engines. Moreover, the censorship practices of the search engines entail the use of broad self-censorship rather narrow interpretations of specific legal notices.

The movement toward greater transparency and accountability must not remain stagnant or in decline. Independent monitoring is required to empirically establish levels of search engine censorship and evaluate search engine companies' compliance with public pledges regarding commitments to transparency, accountability and human rights. This helps prevent backsliding on the part of search engine companies as well as ameliorate any misleading charges levied against them. It also allows companies to access information concerning their competitors' practices that would not otherwise be revealed. An accurate account of search engine censorship is a step toward demystifying and exposing China's Internet censorship policies.

Appendix A

In attempting to develop an automated system that can reasonably compare the search engines some additional methods which would be well suited to one search engine but for which comparable data could not be generated from the others have been delegated to separate search engine-specific projects. As a result of the focus on comparability the methods outlined below not only build upon existing research in this area but also can explain some of the anomalies previously identified. After reviewing previous reports by Reporters Without Borders and Human Rights Watch the methods used to provide an accurate, automated comparison between the search engines are described in detail.

Previous Research

In June 2006, Reporters Without Borders (RSF) conducted a comparison of the four search engines Google, Microsoft, Yahoo! and Baidu (later updated to also include Sohu and Sina) by entering key words into the search engines and analyzing 1) the presence or absence of any results and 2) the content of the results by classifying each returned web site (URL) as either "authorized" or "unauthorized" which presumably refers to whether or not the source is controlled by or supports the government of China or whether it contains critical, alternative information.[42] While this report is an innovative attempt and comparison it suffers from methodological issues that affect the accuracy of the results.

The top ten results from the various search engines were analyzed based on their content, not on whether a web site had been censored (de-listed/removed) from the results set. While the removal of censored sites will likely affect the combination of "authorized" vs. "unauthorized" sources, it does not tell us what sites are censored or if the "unauthorized" sites are not censored but just do not appear in the top ten results. Since localized search engines often algorithmically privilege sites in the local language, ending in the country's domain suffix (e.g. .cn) and possibly even being hosted within the country it affects where foreign hosted "unauthorized" content appears in the result set. Thus an "unauthorized" site may not appear in the top ten results of the localized search engine even though it does in the uncensored version. Instead, the site may appear further down in the rankings.

However, the testing of the search engines did not account for China's national filtering system, often labeled the Great Firewall of China (GFW). Consequently, the results concerning "no results" and "no results + user banned" should actually be seen in reverse. Since Yahoo! and Baidu are physically located in China the search queries made by RSF were filtered by the GFW on their way to the Yahoo! and Baidu servers. If the same search were conducted from China, the search queries would not pass through the GFW and would not be filtered. The search queries RSF made to Google and Microsoft did not pass through the GFW because those servers are not located in China and therefore

[42] Reporters Without Borders. (2006). Test of filtering by Sohu and Sina search engines following upgrade. Retrieved May 22 2008, from http://www.rsf.org/article.php3?id_article=18015

results were always returned.[43] However, had those same search queries been made from China to servers hosted outside China they would have been filtered by the GFW and would have been designated "no results" and "no results + user banned". The failure to account for the GFW prevented RSF from accurately interrogating the filtering of the search engines because a distinction was not made between filtering by the search engines and filtering by the GFW. If the tests had been conducted from inside, rather than outside of China, the report would have captured the behaviour experienced by users in China who are censored by both the GFW and the search engines and perhaps are agnostic about which one is doing the censoring since the result is the same: censorship.

In August 2006, Human Rights Watch (HRW) released an impressive and detailed comparison of Google, Microsoft, Yahoo! and Baidu.[44] Two approaches were used in this report: the first focused on identifying censored sites the second on whether or not the result set returned from a search for a specific key word query was censored. The first approach involved using a list of 25 websites and searching for each website in each search engine (using the site: modifier, discussed below, when possible). If a "censorship notification" appeared and there were no results the web site was censored, but the report also noted instances in which the message appeared but some partial results appeared as well. In other cases, there were no results and since there was also no censorship notification (or a censorship notification that always appeared and had no relationship with the results) it was suspected that the web site was censored. In this way, HRW was able to determine how many of the 25 sites were censored in each search engine.

HRW tested from both inside and outside of China and was thus able to isolate search engine filtering from that conducted by the GFW of China. HRW notes that queries to Yahoo! from outside China generated errors (as in the RSF study) due to the bi-directional filtering of the GFW (see below). The partially censored results (see "Page Censored" below) can result from at least two reasons. The first is that some search queries automatically trigger the censorship notification regardless of whether the results have been censored or not and second because the filtering algorithms of the search engines are imperfect. Google, for example, does not handle port numbers properly[45] and fails to remove results containing port numbers and Microsoft does not handle domains by their root (domain.com) and therefore sub-domains (www.domain.com or dom.domain.com) may not all be removed. Microsoft also does not properly handle URLs that begin with "https". In such cases partial results may be available despite the search engine's attempts to censor.

Another issue (which is still an issue in the methodology discussed below) concerns search engines that do not primarily censor their results directly. Both Yahoo! and Baidu

[43] Google does maintain servers for google.cn inside China, but when requesting google.cn from outside China users will actually query Google servers outside China.
[44] Human Rights Watch. (2006). Race to the Bottom: Corporate Complicity in Chinese Internet Censorship. *Human Rights Watch.* Eds. R. MacKinnon et al. (18,8 (C)). Retrieved April 12, 2008 from http://www.hrw.org/reports/2006/china0806/
[45] Zittrain, J. and Edelman, B. (2002).Localized Google search result exclusions. *Berkman Center for Internet & Society, Harvard Law School*. Retrieved May 22 2008, from http://cyber.law.harvard.edu/filtering/google/

operate the crawlers that index the Internet from inside China and thus do not index sites that are blocked by the GFW. This reduces the need for the search engines to censor their results, as the index itself is already censored by the GFW. This means that there is not a credible technical way to distinguish between sites that are not indexed and sites that are censored. Another issue is that the GFW is not perfect, and normally censored sites sometimes end up in Yahoo & Baidu's index. There have also been some cases in which Yahoo! has removed indexed sites — those not blocked by the GFW — and used a censorship notification as Google does and Microsoft did previously. Therefore, for the most part, Yahoo and Baidu do not need to censor their results, because their index is already censored as their crawlers operate from within China and cannot visit blocked sites.

The second approach used by HRW focused on the issue of keyword filtering by search engines. The question is simple enough; will a search for keyword "a" return results "b"? However, the lack of transparency on the part of the search engines makes the answer to this simple question difficult. HRW used a list of 25 keywords to query the search engines and inferred possible censorship by comparing the results from censored China-specific versions of Google, Microsoft and Yahoo! with their US counter-parts as well as noting the appearance of a censorship message. (Baidu had no such counterpart at the time, but perhaps Baidu Japan can now be used for this purpose.)

Comparing result sets can be problematic because of the algorithmically determined rank of the results. What appears on page one in the top ten results in google.com may appear on page twenty-five in google.cn. In the case of Yahoo! and Baidu GFW-censored sites are not indexed at all and so will never appear no mater what one searches for. Another method is to use the difference in the estimated page count as an indicator of censored results. But the estimated page counts can vary considerably between servers and language/region-specific versions. Microsoft, for example, returns very few Chinese language results in their default English language search engine making comparison virtually impossible. As noted by HRW, even the presence of the censorship notification may not be reliable. In some cases the censorship notification will appear based on the keywords in the query not on the results returned. (One can restrict the results to a non-existent site and still get the censorship message.) In other cases, it has nothing to do with what was used as a query, a non-politically sensitive term, for example) but the censorship message appears because a URL has been removed/de-listed. In other cases, some keyword queries return results and a censor message not because results have been removed but because results are only returned from a set of "white listed" sites.[46] Compounding the problem, the censorship message appears to be page specific (at least in the case of Google). That is, if one searches for keyword "x" and gets back ten results there may be no censorship message, but when one click on "Page 2" and gets results 11-20 which do contain a censored site the censorship notification will appear. (Therefore, if one sets the preferences to retrieve 100 results one may be more likely to encounter the censorship notification than if restricted to 10 results).

[46] See, http://www.nartv.org/2006/06/21/keywords-googlecn/

HRW accounted for such variance through manually checking results in addition to the estimated page count comparisons and the presence of a censorship notification. Not only does this involve extensive manual labour but also an expertise in analyzing the content for political significance. For example, HRW manually assessed and compared the first three pages of search results for Yahoo and Yahoo China. HRW's efforts in this regard stand out as an example of the quality needed for this line of research.

Methodology

The technical component of the Search Monitor Project currently contains two related but separate parts. The first focuses on a generalized comparison between the China-specific search engines of Google, Microsoft, Yahoo! and the domestic Chinese search engine, Baidu. It compares the frequency of censored web sites in relation to key words that are used as queries in each of the search engines. It also compiles and compares censored web sites across all the engines. The second focuses on comparisons between the Chinese-language "global" versions of Google and Yahoo! and their special censored China-specific versions. While the core testing methods are the same, the latter contains some additional elements that allow for a more fine grained analysis. These will be noted below when appropriate.

Generating a URL Set

A set of sixty keywords have been selected covering the broad topical categories of censorship circumvention, the Falun Gong movement, political sensitivities and social taboos.

Search queries in an uncensored search engine (the Chinese language versions of Google) are used to generate lists of sites that are checked in censored search engines.

A query term, such as "人权" (human rights), is used to retrieve results from an "uncensored" search engine, such as google.com.

The websites from the "uncensored" results are parsed to retrieve the domain (including sub-domains).

A list of URL results (ten) is retrieved. A URL, such as http://www.hrw.org/chinese/, is shortened to its domain, www.hrw.org

Each domain name is checked in each censored search engine.

Determining a Censored site

Domains are checked in the censored search engines using the "site:" modifier. The "site:" modifier restricts the results set to pages of a specific host name.

> "site:www.hrw.org" (without the quotes) is used as a search term in censored search engines to restrict the results to only those from the web site www.hrw.org

In cases where the censored search engine being tested displays a special message indicating that results have been censored, a "censor message", that relates to the specific search query, domains that produce no results when queried with the "site:" modifier and contain a censor message are labeled as "Censored" while domains that return some results but contain a "censor message" are labeled as "Page Censored".

In the cases where there is no censor message, or the censorship message appears on every page and bears no connection to the results, domains that produce no results when queried with the "site:" modifier are labeled as "Censored."

Depending on the current behaviour of search engines there may be ad hoc additions.

- http://www.google.cn/ - censored = censor message + 0 results, pagecensored = censor message + some results
- http://www.live.com/?mkt=zh-cn - censored = 0 results, pagecensored = results that only contain urls beginning with "https" (no longer a censor message, failure to exclude "https" urls was noted when the censor message was in place and is thus used as "Page Censored")
- http://www.yahoo.cn/ - censored = 0 results, censor message is ignored because it appears on every page, it bears no relation to search results
- http://www.baidu.com/ - censored = 0 results

It is important to note that sites which are simply not indexed by the search engine will appear as "Censored", thus inflating the total amount of censorship attributed to search engines that do not have a censor message that is related to the results. This can be slightly compensated for by looking at the overlap of censored sites among search engines. In addition, since this is a normative project advocating transparency, this should serve as an incentive for search engines to implement a censor message that is related to the results.

The comparisons Google to Google.cn, Yahoo! to Yahoo.cn contain the classifications "Returned" and "Indexed" in addition to "Censored" and "PageCensored". "Returned" refers to URLs from the uncensored search engines that are returned in the result set from the censored search engine. "Indexed" refers to URLs from the result set from the "uncensored" search engine that are not returned in the result set from the censored search engine, but are not censored. Using this method, the top ten results or a query in

Google (Chinese) can be compared with the top ten results of Google China and can be categorized by "Returned", sites that are common to both results sets, "Indexed", sites in the top ten uncensored but not in the top ten of the censored results, and "(Page) Censored", results that are actually censored. In addition, each URL in both result sets is check to see if it is hosted in China or ends in a .cn domain suffix.

The Great Firewall (GFW)

Borrowing a phrase from Richard Clayton, Steven Murdoch and Robert Watson, it is necessary to "ignore" the filtering conducted by China to accurately test levels of censorship by the search engines themselves. As Clayton, Murdoch and Watson reveal, Internet traffic to and from China passes through a filtering system that is bi-directional - it affects both inbound and outbound traffic - which disrupts connections if the presence of particular keywords are detected.[47] Often, China will designate a domain name as "key word" causing the disruption of any request that contains that domain name. This is important as queries directed to search engines hosted in China use the "site:" modifier followed by a domain name.

In order to avoid interference from the China's filtering system, the China-specific versions of Google and MSN, which maintain servers hosted outside of China, are queried from outside of China and the China-specific versions of Yahoo and Baidu, hosted inside China, are queried from inside China.[48]

In order to test from within in China, this project uses TOR exit nodes located within China. TOR is an anonymity system that encrypts the connection between the testing computer located outside of China and the TOR exit node inside China. This ensures that GFW is unable to interfere with the requests made to the search engines inside China.

- The censored search engines, http://www.google.cn/ and http://www.live.com/?mkt=zh-cn are checked from outside of China.
- The censored search engines, http://www.baidu.com/ and http://www.yahoo.cn/ are checked from inside of China.

In addition to affecting how to test each search engine, the location of the search engine to the GFW also affects how the search engines censor. Google and Microsoft, located outside of China, must remove, or de-list, specific sites from the results. Yahoo! and Baidu both operate their search spiders from inside China. The results in a situation where, because of China's gateway filtering, the crawlers that index content for these search engines cannot access sites that China blocks.

[47] Clayton, R., Murdoch, S. and Watson R. (2006, June 28 - June 30). Ignoring the Great Firewall of China. *Paper presented at the 6th Workshop on Privacy Enhancing Technologies*, Cambridge, United Kingdom. Retrieved May 22 2008, from http://www.cl.cam.ac.uk/~rnc1/ignoring.pdf

[48] Google maintains servers for google.cn both outside and inside China.

- 61.135.166.102 - - [08/Feb/2008:08:05:40 -0500] “GET / HTTP/1.1″ 200 12258 “-” “Baiduspider+(+http://www.baidu.com/search/spider.htm)”
- 220.181.38.169 - - [08/Feb/2008:09:04:42 -0500] “GET / HTTP/1.1″ 200 12258 “-” “Baiduspider+(+http://www.baidu.com/search/spider.htm)”
- 60.28.17.38 - - [08/Feb/2008:11:46:31 -0500] “GET / HTTP/1.1″ 200 12258 “-” “Baiduspider+(+http://www.baidu.com/search/spider.htm)”
- 202.160.180.184 - - [07/Feb/2008:16:58:33 -0500] “GET /robots.txt/ HTTP/1.0″ 200 24 “-” “Mozilla/5.0 (compatible; Yahoo! Slurp China; http://misc.yahoo.com.cn/help.html)”
- 202.160.180.96 - - [07/Feb/2008:16:58:35 -0500] “GET / HTTP/1.0″ 200 19068 “-” “Mozilla/5.0 (compatible; Yahoo! Slurp China; http://misc.yahoo.com.cn/help.html)”

Thus Yahoo! rarely has to de-list specific websites; most are just not indexed in the first place. However, this also leads to situations in which sites blocked by China and de-listed by Google and Microsoft are indexed by Yahoo!. The GFW is not 100% effective and occasionally crawlers operating from inside China are able to index a normally blocked site which then appears in their search results.

It is also important to note that sites indexed by the search engines that are blocked by the GFW will still be inaccessible to users in China.