

Bypassing Internet Censorship for News Broadcasters

Karl Kathuria

*British Broadcasting Corporation and Canada Centre for Global Security Studies
Munk School of Global Affairs, University of Toronto*

Abstract

News organizations are often the targets of Internet censorship. This paper will look at two technical considerations for the BBC, based on its distribution of non-English content into countries such as Iran and China, where the news services are permanently unavailable from the official BBC websites: blocking detection and circumvention. This study examines an internal BBC prototype system built in 2010 to detect online censorship of its content, and evaluates potential improvements. It will also review the BBC's use of circumvention tools, and consider the impact and execution of pilot services for Iran and China. Finally, the study will consider the technical delivery of the BBC's news output, and the methods it employs to bypass Internet censorship.

1. Introduction

The British Broadcasting Corporation (BBC) provides programmes and content for radio, television, online, and mobile phones in English and 27 other languages. There is currently an increased focus on delivery of its services online, as the amount of radio content has reduced, both in terms of hours of output and infrastructure for shortwave delivery. However, some of the services that are considered strategically important are for countries in which BBC news content has been blocked, either short term or persistently.

This paper looks at two technical considerations for the BBC, based on its distribution of non-English content into countries such as Iran and China, where news services are permanently unavailable from the official BBC websites: blocking detection and circumvention. It will then consider other ways in which content can be delivered where BBC web sites are inaccessible, and how the BBC can continue to reach its audience when its services are blocked.

2.1. Detection: The Brief

When the BBC knows that one of its web sites has been suddenly made unavailable due to a blocking event, it can put into place processes for dealing with that block. However, it first needs to gain an awareness of when and why its services are being blocked. To address this requirement, the BBC developed a software prototype — Geostats — to detect blocking events.

The development of Geostats was motivated by two factors. First, during the first half of 2010, there were several “false alarms,” where audience members had written to the BBC language services to report that con-

tent was being blocked in their country. When these claims were investigated, it was found that the inaccessibility was not due to filtering, but likely the result of various network conditions and outages that made content unavailable for short periods of time, or in some cases, improperly configured computers. Investigating each of these possibilities was taking up time and resources for the team responsible for content distribution.

Second, conversations with Google revealed that their Transparency Report [1] was nearing completion. What was learned from Google would help in devising a project plan for the BBC to create its own software to detect censorship. The Geostats project was set up using technical experts from BBC World Service, who developed a system based on five key requirements:

1. Interpret traffic data from two sources: Livestats, the BBC's own “web bug”-based software used publicly to display the current Most Read / Most Emailed stories on bbc.com/news and World Service language web sites in near real-time; and server logs from the streaming media provider, showing technical details regarding the serving of streaming media.
2. Separate the traffic data by country.
3. Normalise the data in such a way that the system would generally report traffic along the zero-line of a graph regardless of the time and day. Extremes could then be attributed to either major news events or possible blocking of content. Any time the data was +/-60% of ‘normal’, an alert state would be reported.
4. Structure the system in such a way that it could query external data sources for more information when an alert state was reached.

5. Provide multiple views of the data, so that it could be used both for technical analysis and management summary

2.2. Detection: The Implementation

There are three main processes behind the Geostats implementation: database import, building the behaviour model, and displaying the results.

Two sets of data were collected. First, Livestats data was received hourly via an API call to the system, and put into a database table. Data captured included country code, timestamp, number of hits and number of unique IP addresses.

BBC streaming media is hosted by Akamai, the Content Delivery Network (CDN). Log data is collected hourly, and contains information about every piece of streaming media served by the BBC language sites, including the IP address of every computer accessing the audio and video. Approximately 30GB of uncompressed log data is collected daily.

The data from Akamai is then taken per hour per log file and put into a similar table, with a country code, timestamp and number of hits.

For both database tables in the prototype, the country codes were found with geographical IP lookups using software from Maxmind (maxmind.com). The next step is to normalise the data, and to identify the expected traffic at any given hour in the day. Scripts are scheduled to run hourly (using crontab), converting the new data into JavaScript Object Notation (JSON) objects. This data is presented using PHP and Open Flash Charts (OFC).

Figure 1 shows the shape of processed data over a 3-day period for both collection methods, showing content served to Vietnam. The yellow line shows traffic from bbc.co.uk, and the grey line shows streaming media traffic served via Akamai. Every day, there is a morning, afternoon, and evening peak for traffic—a pattern that is repeated throughout the week, albeit with lower traffic levels at weekends.

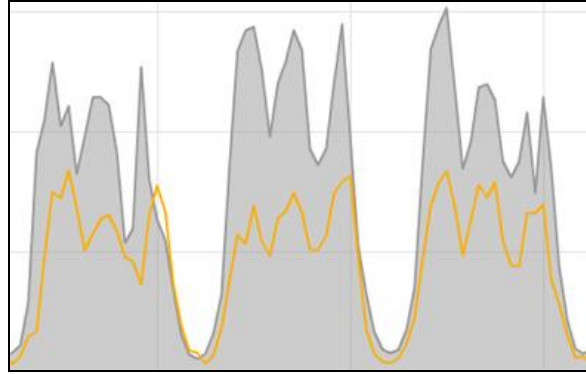


Figure 1: Shape of daily traffic to Vietnam over 3 days

The simple calculation used to create a ‘normal’ value was based on an average for any given hour and day of the week, built up over the life of the system. Therefore, the expected level of traffic for Monday at 14:00 would be defined by the average of all previous traffic levels for any given Monday at 14:00.

The resulting average is inserted into another database table for comparison against current levels. The fields are: hits; country code; hour of day (0-23); day of week (0-6). This method of averaging the content is imperfect, as it only compares current traffic against a limited set of other data, but provided an acceptable starting point. Future iterations of the system would introduce traffic comparisons by the hour and day, as well as similar days of the week.

At this point, a graph can be created to show the level of traffic compared to its expected value, and for this deviation to be plotted around the zero-line, the level defined as “normal”. After a month of data collection, the graphs were showing a large degree of variance, based mainly on the short period of time for which data had been collected.

At this stage, the alert threshold was set to a deviation of +/- 60% to allow for the small amount of normalising data. If traffic at any given hour is found to be outside of this range, an alert state is generated. For example, the red lines displayed in figure 2 indicate a large dip in traffic for certain hours in Pakistan.

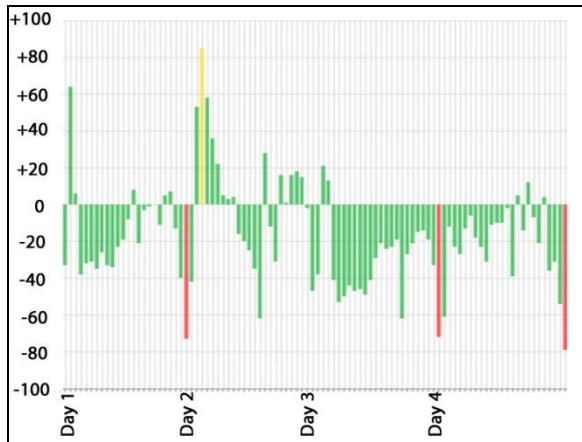


Figure 2: Four days of traffic deviation in Pakistan

In the prototype Geostats system, an alert is displayed to the user when hovering over any red line displayed on the graph. The concept behind this alert is that the system would retrieve information from an external source with an API call. The “click to see more” message reports any information related to that alert to better inform users why the alert is being displayed (see figure 3). The system was set to retrieve any news reports from the OpenNet Initiative’s RSS feed related to the country being analysed.¹

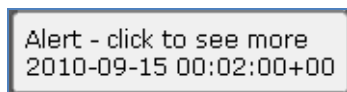


Figure 3: Geostats alert message

The final component of Geostats is a “calendar view,” designed to be easily interpreted by non-technical users. This view took the same information, but presented a daily alert for each country (see figure 4). A country would show as having a problem if there was an hourly alert state for more than four hours in a 24-hour period.



Figure 4: Calendar view

¹ For example the ONI RSS feed for Pakistan is <http://opennet.net/taxonomy/term/41/feed>

2.3 Detection: Lessons Learned and Potential Improvements

Geostats was developed as a prototype with the intention to improve it based on lessons learned during development. Areas for improvement were identified at the three stages of the system:

Database import:

- Log files are up to 24 hours behind real-time, which means the system cannot report effectively on a full set of live data.
- The size of the log files requires the system be optimized for performance.
- Malformed lines in log files are common and can skew the data. Refining the parsing methods could provide a more robust system.

Building the behaviour model

- Building the relationships places a huge load on the server, both in terms of processing power and database calls.
- The geographic lookup is imperfect, and also can take a long time. The ideal solution would give more granular views of outages, separating not just by country but by region and ISP. This implementation would take additional processing power, and also have a financial impact.
- The normalisation process could be improved. One approach is to do more calculations to assess any given hour against the previous hour, previous day, and every seventh previous day in the system.

Displaying the results

- The graphs and calendar view are all created at the time of request, with queries being made to the database.
- The processing power required to deal with this request resulted in processing times of up to 5 minutes for the system to return results. Some form of caching could be introduced to improve this.

Aside from these issues, there are two main challenges for the system: handling alerts and providing API access. For handling alerts external APIs from reliable data sources can be queried, with the BBC able to provide three variables: country, date of alert, time of alert. The structure of the information that is returned from this query will need to be defined so that it can be imported into the BBC’s Geostats system. The second challenge is to provide API access to Geostats for similar systems. A table in the database could be created, with an API to expose the country code, timestamp and percentage variance from ‘normal.’

3.1 Circumventing Blocks

Around the time of the Iranian elections in June 2009, the BBC Persian web site was suddenly blocked in Iran. The site had been blocked on previous occasions [2], but this was the first block since the BBC Persian satellite television station was launched several months earlier. BBC Persian TV was also blocked, prompting a blog piece from the BBC's Director of Global News claiming that such actions were "against international treaties on satellite communications" [3].

During the block, the BBC noticed a huge increase in traffic to the BBC Persian TV Internet live stream of more than four times its usual levels. Geographical IP lookups of this traffic showed that the majority of streaming was from inside Iran. The BBC began to find ways to let the audience know that it could still access the streaming media.

Psiphon Inc., a privately held Canadian corporation that develops circumvention technologies, provided proxy servers configured with `bbcpersian.com` as the first page that would be seen after logging in. The BBC used `bit.ly`, the URL shortening service, to set up links to Psiphon servers and to the direct media streams, and used all available channels to promote the availability of BBC content: Email newsletters, Twitter, Facebook and on-air promotions. The use of `bit.ly` allowed the BBC to look at near real-time statistics showing what country the shortcut URL was being accessed in. A separate web page was created to carry the BBC Persian TV stream through a non-BBC HTML player which was not being blocked in Iran.

Having evaluated various tools prior to this incident, Psiphon was considered the most suitable for the organisation, as the BBC had no funds for development of new tools or additional technical support. Of the available "off-the-shelf" solutions, Psiphon had these advantages:

- Usability: The BBC would direct its audience to the tool, but not be responsible for hosting it.
- No executable code: The BBC was not willing to provide software that had to be installed on the user's PC, therefore a web-based proxy was chosen.
- Appropriate information: The user of the circumvention solution would need an appropriate set of clear terms and conditions for use.
- Hosting environment: The hosting of the solution had to be in a secure environment, provided by a trusted corporate entity, and not peer-to-peer based.

- Exit strategy: If a web-based solution was compromised or if the BBC needed to withdraw from providing the service, it had to be able to do so.

3.2 Adding a Service

While the Psiphon service for BBC Persian grew organically from the sudden blocking of the web site, the service for BBC Chinese was introduced at a time when it was needed, and with the involvement of its production team. At the end of September 2010, the BBC Chinese service started to publicise Psiphon nodes, with the message "If you are having trouble accessing our site in China, please try [node URL]."

Three nodes were set up by the Psiphon team, with their URLs propagated through different channels by the BBC: one through email newsletters; one through the BBC Chinese Twitter account, and one promoted on air and in direct email contact with individuals.

The frequency of propagation differed for each of these channels. A message accompanied every radio broadcast (three times per day, with additional trails featuring the promotion), email newsletters were sent daily, and Tweets were sent on an ad hoc basis.

Figure 5 shows the number of pages viewed through the BBC Chinese Psiphon service in comparison to BBC Persian during the first eight weeks. The active promotion from BBC Chinese helped the service grow rapidly so that the number of pages served was matching those served by BBC Persian's Psiphon implementation.

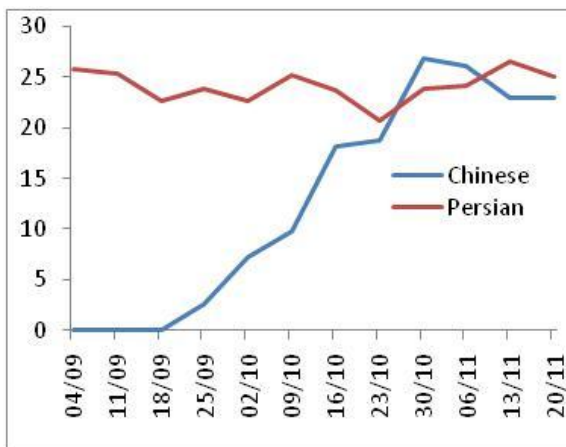


Figure 5 Pages Served (000s) to BBC Psiphon servers (date format dd/mm)

3.3 Dealing with a Short-term Block

Shortly after the BBC Chinese service started propagating the availability of Psiphon nodes, a news event led to additional blocking of BBC content inside China.

The Nobel Peace Prize ceremony to honour Liu Xiaobo was due to take place in Oslo on December 10, 2010.

On the morning of December 9, 2010, the BBC published a story with the headline “Nobel: China blocks foreign web sites ahead of ceremony” which reported on the blocking in China of a number of sites, also including Norwegian broadcaster NRK [4]. BBC sites that were blocked included the English news site, and bbcukchina.com, an educational site with content about life and culture in the UK. The block also caused an increase in undeliverable email newsletters for BBC Chinese.

The BBC worked with Psiphon to offer two extra proxy servers to bring visitors directly to the BBC News site in English. With the service already being blocked, this was considered sufficient to cover the likely audience. The first was promoted via Facebook and email newsletters during the day, while the second node was brought online at night via Twitter promotion.

Plans were made to provide a video feed of the live ceremony that would be BBC-branded, but not hosted on a BBC site. A bit.ly URL was created for this feed. Editorial teams were instructed to only propagate that message one hour before the start of the ceremony to minimise the likelihood of it being blocked.

Of the two separate nodes brought online for the BBC News site in English, one was blocked almost immediately, while the other was available throughout the weekend. On the day of the ceremony, there were 387 logins to this server. While this is a small number of people, it was considered a success in that there was no direct way of promoting the availability of the service due to the sudden blocking of the BBC News site.

A live stream of the ceremony was also created on a non-BBC branded page, with bit.ly URLs promoted one hour before the ceremony. While there was no expectation of a huge audience for this stream, data from bit.ly shows that there were 4,236 clicks to that URL that day, with around 50% from China. This traffic accounted for about one-third of the total visitors to that stream, including those accessing it from unfiltered locations.

On December 13, the BBC news site in English was again available in China, which was reported with the conclusion, “China appears to have done what it could

to stop unfiltered news of the event reaching its own people”[5].

4.1 Improvements to BBC Web Sites

BBC sites are not always rendered correctly using Psiphon. Freedom House noted this issue in their review of circumvention software, pointing out that web-based proxies have an “inability to properly translate flash and some other forms of dynamic content” [6].

There are two areas of concern for the BBC pages:

- Page structure: There is a heavy reliance on JavaScript, particularly on the index pages of language news sites such as bbcpersian.com.
- Audio and video: Audio and video on the BBC’s language web sites is served using Flash Player or Windows Media Player, which will not be accessible through a web proxy.

The BBC could work with Psiphon to improve content caching and JavaScript delivery. This approach has proven to be successful with Voice of America sites, which render properly using Psiphon. However, many modules on the BBC language news sites are used to render sites covering all of the BBC’s entertainment output as well as its news. Changes made in this environment to support serving content through web proxies would likely be considered low priority. An alternative is to consider changes that will only have an effect on the language news sites, such as modifying the CSS.

The BBC is producing some audio and video content in formats that can be delivered via HTTP rather than RTMP. This standards-based approach allows content to be consumed on newer mobile devices, specifically those with support for HTML5, and will make BBC content more accessible through Psiphon and other web proxy tools. BBC sites with audio and video available over HTTP have been successfully tested over Psiphon.

4.2 Use of CDNs

The BBC’s international news services have distributed audio and video content through Akamai since 2003, and URLs for the streaming media do not include bbc.co.uk. Instead, Akamai’s edge network handles the stream request so that the server providing content to the end user is the one nearest to that computer on the network [7]. As the servers are not specific to the BBC, it makes URL and IP-based blocking more difficult than targeting *.bbc.co.uk. There have been reports of individual machines on Akamai’s network being blocked, which has led to “thousands of web sites” also being

made unavailable [8]. The effects of such filtering are much wider than a blocking of an individual site.

The BBC uses multiple CDNs to provide its streaming media. Live streams are available to mobile devices through non-BBC URLs such as iPhone video streams and SHOUTcast MP3 streams, while news clips for mobile devices are provided in discrete indices on the mobile version of the web site, with device detection and media hosting through a CDN. Clips are available through HTTP, so web proxy users can consume BBC language news videos through the mobile sites. However, the video index is only provided as a chronological list and offers no editorial ordering of content.

4.3 RSS and Syndication

For most BBC language news sites, only headline and summary text is available for public syndication, with full text offers restricted to commercial deals. However, for BBC Persian, there is a ‘full-text’ feed available publicly, as the BBC recognises the difficulty of serving content into Iran. This format, and its availability in multiple blogs and web sites, makes it difficult for Iranian authorities to find and block BBC content.

RSS feeds are also used for distributing content via iPhone apps that have been created by an enthusiast of the BBC Persian site. The developer uses the RSS feeds to recreate web content inside the app, and also links to the live audio and TV streams, and on-demand video files, carrying web content over HTTPS. The Chinese app is available in China’s app store, and information from the developer suggests that downloads inside China count for over 50% of the total for the app.

Audio and video content is also syndicated, and one of the BBC’s partners for its language video content is YouTube. YouTube channels exist for BBC Persian and BBC Chinese, as well as several other BBC languages. Statistics from YouTube show that when YouTube itself is available, BBC videos can be viewed through the site inside both Iran and China.

5. Conclusion

As the BBC focuses more of its efforts on online platforms for news delivery, the need to detect and circumvent blocks becomes increasingly important. The pilot circumvention service and research into the distribution and blocking of its content has increased the understanding of its online environment. This project has highlighted improvements that can be made and challenges that need to be overcome.

The BBC has recognised that bypassing censorship involves more than merely providing software that circumvents blocks of its content. There will be times when a complete throttling of internet traffic, such as recently experienced in Tunisia and Egypt, will make its content distribution more challenging, through circumvention tools as well as directly. The way its services are hosted will help BBC distribute its content into countries where it may otherwise be unavailable. Alternative methods of consuming news need to be promoted and constantly updated, and communicated to its audience through whatever channels are available.

There is evidence through propagation of circumvention tools and multiple distribution channels that the BBC is able to reach people who want to access BBC news. The broadcaster needs to know when it is being censored, and that its role at these times is to direct its audience to its content, wherever it is available.

6. References

- [1] Google, “Transparency Report,” <http://www.google.com/transparencyreport/traffic/>.
- [2] “Iran Blocks BBC Persian Web Site,” BBC News, January 24, 2006, http://news.bbc.co.uk/1/hi/world/middle_east/4644398.stm.
- [3] Peter Horrocks, “Stop the Blocking Now,” June 14, 2009, BBC News http://www.bbc.co.uk/blogs/theeditors/2009/06/stop_the_blocking_now.html.
- [4] “Nobel: China Blocks Foreign Websites Ahead of Ceremony,” BBC News, December 9, 2010, <http://www.bbc.co.uk/news/world-asia-pacific-11962520>.
- [5] Michael Bristow, “BBC Website Unblocked in China,” BBC News, December 13, 2010, <http://www.bbc.co.uk/news/world-asia-pacific-11981014>.
- [6] Freedom House, *Leaping Over the Firewall: A Review of Censorship Circumvention Tools*, April 11, 2011, <http://www.freedomhouse.org/template.cfm?page=383&report=97>.
- [7] Akamai, “Streaming” <http://www.akamai.com/html/technology/products/streaming.html>.
- [8] “Akamai blocked in china cause thousands of website also been blocked,” Twit Browser, <http://twitbrowser.net/33>.