This study investigates prevailing restrictions on Internet diffusion through blocking of websites across five categories in diverse country domains. A website block finder (WBF) application is used to provide a snapshot view on Internet filtering imposed through government regulations across ten countries, namely Canada, China, France, Germany, Iran, New Zealand, Saudi Arabia, Turkey, UAE and USA. The five categories selected for this study are religion, political, news channels, entertainment and social networking. The WBF performs a domain name server lookup and searches for websites which are identified by their IP addresses by following related linked URLs to inform whether websites belonging to that category have been blocked or not within that country’s domain. Findings reveal that no blocking restrictions existed among the identified categories in four countries (Canada, New Zealand, Turkey and USA), but varying degrees of restrictions have been applied in the remaining six countries (China, France, Germany, Iran, Saudi Arabia and UAE) at the time the WBF application was run. The study provides a snapshot view on website restrictions within the selected categories and raises awareness on prevailing regulations imposed on Internet diffusion in the ever-increasing complex global environment.

Keywords: information filtering, website blocking.
1 INTRODUCTION

Societies today are influenced by the virtual world of the Internet in which different cultural groups interact over blogs and social networking sites which have become everyday platforms to connect people and disseminate information across national borders. The evolving Web has advanced literature in sciences, arts, languages and businesses as open discussion forums across different regions have given rise to new thinking and commonality in terminologies. For example, words like google (search) or unfriend (remove a friend from social networking sites) have become commonplace words, and have been added in English dictionaries. However, the evolving Web also has negative influences and has raised concerns involving security, piracy and privacy issues as some websites have been created specifically for on-line scams, illegal file sharing, phishing and spamming amongst others. The Web is also considered to have a negative influence if user views expressed on open discussion forums are against government’s political views or country’s religious sentiments or other such types of patterned beliefs. Viewing both sides – positive and negative – of the debate, the Internet has been compared to a sword with two edges (Hamade, 2008).

To address some of the negative influences, many directives are enforced for policing the diffusion and adoption of Internet, and the area of Internet regulation is of growing importance. This implies that Internet traffic is monitored by external agencies and certain websites may be censored due to their inappropriateness within some social, political or religious context. The censorships could occur at various levels: national, organizational and individual. At national level, governments restrict some access areas based upon their political views, liberalisation policies, religious sentiments, or to prevent online imaging of sensitive areas in their country (Hamade, 2008, Cankorel & Aryani, 2009, Opnet, 2009); at organisational level, organisations restrict access for organisational privacy reasons or if they consider social leisurely activities such as web browsing/ chatting lowers productivity (Chou, Sinha & Zhao, 2010); and individual level restrictions imply filtering tools used by parents to protect their children from some of the harmful Internet content (Hamade, 2008, Opnet, 2009).

In this study, we investigate website filtering on the public Internet at the national level to provide a snapshot view on government applied censorships across ten countries, namely Canada, China, France, Germany, Iran, New Zealand, Saudi Arabia, Turkey, UAE and USA. Accordingly, a software referred to as website block finder (WBF) application has been designed to measure blocking of websites across five categories namely, religion, political, news channels, entertainment and social networking. Moreover, because of the different beliefs and multitude of perceptions on religions practiced worldwide, the religion category encompasses five distinct religions and the atheist category which promotes views of non believers. We begin this paper by discussing the commonly used methods for Internet filtering and also describe some methods used to bypass filters imposed at the national level. Next, we discuss the research design of WBF used for conducting our study and data collection methods used. The findings are presented and discussed as new insights are shared on current Internet regulations.

2 INTERNET FILTERING METHODS

Governments adopt policies for regulating Internet adoption and diffusion, in which they outline the rules, procedures and controls used by Internet Service Providers (ISP) for managing information flow within their country. One such policy refers to Internet filtering, where controls are enforced to conform to social, religious or political standards on what is considered lawful and should be allowed, or what is considered unlawful and should be blocked or filtered within that country. Thus the terms Internet filtering and website blocking are used synonymously to imply prevention of access to certain websites which are considered inappropriate for public access.

Each website hosted on a server can be identified through a distinct number which is assigned to every online device called IP number. The IP number is a unique 12 digit number including digits from 0 to 255 and is divided into three digit segments (e.g. 123.124.255.101). Furthermore many
websites may have a one domain name but many IP addresses. For example, Google has linked several IP numbers such as 74.125.53.99 (google.com), 74.125.155.99 (google.co.nz) and 72.14.213.99 (google.co.au) amongst many others. This helps in streamlining the Internet traffic so that the flow of information is evenly distributed on networks; also ensuring search results are customised to project latest events and happenings for that geographical area (e.g., the top listings of Google search result from google.co.nz would be more relevant to the public of New Zealand than to the public of any other country). Hence, all websites have a domain name and some associated IP number(s). Once a website is differentiated from another through its IP number(s), it can easily be blocked if it has been identified as a candidate containing objectionable content. However such IP blocking methods are referred to as “blunt filtering methods” since they block out even acceptable websites simply because these are hosted on the same server as a restricted site (Opnet, 2009).

The filtering mechanisms are mostly applied to ISP gateways through URL-blocking via black-lists and white-lists. The black-list contains a list of URLs considered inappropriate and must be blocked, while the white-list contains a list of permissible sites (Polpinji, et al., 2008). However URL-blocking through black-lists and white-lists has overheads since the lists need to be updated frequently as new websites are emerging daily. Moreover, governments mostly keep black and white lists confidential (Hamade, 2008) so as not to publicise the fact that they have restricted the Internet access. The ISPs are advised to redirect network traffic by rewriting network address books to point to some other website or to return block-pages instead of the requested page (Hamade, 2008). The block-pages may either acknowledge the fact that the website has been blocked or the block-page may be disguised as an error message. ISPs also maintain log files of websites accessed by users, and hence each user who tried to access a restricted site can be identified. Thus, web un-blockers advise users to never bypass blocking after being stopped by an ISP, since ISP administrators can find user preferences through the log files. This information may have ramifications on the user’s privacy and confidentiality depending upon how governments view such infringes.

Another approach used in regulating access to websites is through keyword filtering (Du, Reihaneni, and Siusilo, 2002; Lee, Hui, Cheuk & Fong, 2002). In this method lists of keywords are marked as objectionable, and websites containing any of these words are identified and blocked. However, keyword filtering blocks all websites containing these words, although we know that the meaning of words cannot be taken in isolation as the meaning of the word changes with the subject context (Polpinji et al., 2008). For instance the word ‘gay’ can imply a person being happy or could describe a person’s sexual orientation within different subject contexts. Hence content-based analysis technique is more suitable than just keyword filtering (Lee et al., 2002).

Search robots or crawlers are used to locate websites containing sensitive or objectionable words. The crawler is bootstrapped with a list of URLs, called “seeds”, which are the access nodes from where the web search crawl is initiated. The crawler (or spider) “iteratively collects information from the web, downloading pages and following the linked URLs” based upon some harvesting strategy (Gomes & Silva, 2006, p. 1) such as searching for words considered sensitive within that country domain. The harvesting strategy by spiders yields a list of websites, which are evaluated by regulating agencies for appropriateness within their country context. Any website not considered to be appropriate, is then added to the black-list and banned by ISP gateways.

### 3 HOW TO BYPASS INTERNET BLOCKING

Though Internet censorship is applied at different levels, it is possible to bypass the blocking, but this could involve taking risks of breaking the law (Crowcroft, 2007). Some methods to bypass blocking are now discussed.

One method involves connecting to a computer outside the country domain and using it as a proxy, to access the black list. Since a country cannot block websites outside its borders, this method is popular with citizens wanting unrestricted access to the web. Psiphon software tool developed by University of Toronto's Citizen Lab can tunnel through social networks, in which users in uncensored country can download the program, and transform their computer into an access point or proxy. Now users
belonging to censored places can be given a unique web address, login and password, to enable them browse the web freely through their encrypted connection to the proxy server. Another method to bypass the website blocks is through use of cgi-scripts for encrypting URLs. Software programs known as anonymisers use layers of encryption algorithms to provide anonymous connections using complex code schemes, website aliases and third party mirrored sites (Argyrakis, Gritzalis & Kioulafas, 2003). The encrypted message from the anonymiser is not detected as a black-listed URL. Many commercial and not for profit organisations provide services for URL encryption and other similar requirements.

Commercial virtual private networks (VPN) are deployed widely, which enable secure data transmission across the Internet. Encryptions are applied to preserve authentication and privacy through secure shell network protocols (ssh/sshd). Further point-to-point protocol (pppd) establishes a working connection between the VPN servers and user workstations (Alex & Fei, 2008). The pppd does not appear in the ISP gateway records and hides user activities from being detected. The user sitting at a local machine is transformed into a remote machine user with a virtual IP address and can freely access websites. Moreover, VPNs have many servers in different countries allowing users to shift their IP between servers adding to user anonymity. So a VPN is an attractive option for accessing restricted websites, more so through public unsecured WiFi networks. Unsecured WiFi networks can be used undetected by anyone who may not have subscribed to the VPN connection as long as they are within the transmission range. However, VPNs could impose security threats to users from each other, as the network does not know what traffic is flowing within the VPN and users could infect each other’s computers in the network (Alex & Fei, 2008).

Another technique to bypass regulatory web blockers and to view web pages is through email services. For instance, email services such as www.web2mail.com provide content of requested webpage via emails. Users send email to www@web2mail.com with the requested web address in the subject line, and receive online content via email. Moreover, if users frequently request particular web pages, these services register the requested pages which are then updated frequently. However, researchers warn that these technologies are slippery slopes and nothing is really anonymous in the digital world. They warn that users leave some data trail which can be later reassembled by digital forensics into a detailed account of individual user’s activities (Aaron, 2009). Websites offering services for bypassing web blocks also warn users to be aware of the risks they are exposed to by using their service.

4 RESEARCH DESIGN METHODS

This section discusses the research design of the WBF and methods used in detection of blocked websites. The WBF has been constructed using the Visual Studio C++ MFC. The blocking methods explained in the previous section have been used in the construction of the WBF.

First, lists of popular or top ranked websites were identified using Google search engine. A search was conducted on words like “Islam Sunni”, “Islam Shia”, “Christianity”, “World News”, and “Entertainment Movies” amongst others. The top ranked results obtained from the search were double checked by authors to ensure they contain information of that particular category. In this manner ten websites from each category were listed randomly. This process ensured that the list was not selective or biased towards any country’s current social, religious or political beliefs. Next, domain name server lookup was performed to locate IPs through methods such as reverse DNS lookup methods, command prompts (such as “ping” or “traceroute”), and through use of “whois” tools. “Whois” is a service in which many websites have registered their details, such as location, IP address, and purpose (http://whois.domaintools.com). A list containing domain name and IP address in the five categories was created and is referred as pick-list. Finally, the html content for each website listed in the pick-list was added to the pick-list detail file.

Figure 1 shows the WBF design for checking whether websites in the pick-list are blocked or not. The WBF application is developed in Visual Studio C++ MFC platform and compiled into an executable (WBF.exe) file.
The WBF first performs a domain name server lookup and pings each URL address in the pick-list sequentially. The IP address of each pinged website is compared with the IP address in the pick-list. However, as has been stated earlier, some domain names such as search engines have more than one associated IP address. Web crawlers on the Internet crawl through related linked URLs, and check if any of the attributed IP addresses match those specified in the pick-list. If the IP address does not match, this could be due to network traffic being directed to some other website (e.g. block-page). Accordingly if no match occurs, the domain name is added in the blocked-list. But, if the IP address matches the pick-list, then the HTML content of the visited website is compared with the HTML content stored for that URL address in the pick-list. If the HTML content does not match, this also means that some percentage of blocking exists for the webpage, and this domain name is added to the blocked-list.

Thus, the blocked-list or the output file contains a list of websites which may have either been totally blocked or may have had some objectionable content filtered out from the original website. In this manner, the WBF gives a snapshot view of website blocking in the five chosen categories.

5 DATA COLLECTION METHODS

This section explains how data has been collected from the ten countries selected to participate in this study. Though blocking of websites is also based upon the local language, this study has considered only English language-based websites in the pick-list. Use of a common language (English) further ensures consistency in data collection across the sample countries.

The authors have access to a VPN having servers in five countries namely, Canada, France, Germany, New Zealand and USA, which has been used to test the WBF in these countries. When a connection is made to a server in a particular country domain, the user can access only the websites allowed within that country. Thus, connections were made to different servers by the authors and executable file (WBF.exe) run to output the blocked-list file for that particular country. For the remaining five countries (China, Iran, Saudi Arabia, Turkey and UAE), the study identified participants residing in these countries. Overall, participants belonging to eleven countries were asked to participate in the study but only five responded. Each of them was asked to run the WBF executable file. Every participant was emailed the executable file, which was run in that country domain. The output file containing the blocked-list was later emailed to the authors by the participants.

6 FINDINGS AND DISCUSSION

The selection criteria of websites in the pick-list belong to five categories, namely religion, political, news, entertainment and social networking are explained next:
The religion category identified five distinct religions (Islam Shia, Islam Sunni, Christianity, Judaism and Bahai) since these are some of the main religions practiced worldwide. Furthermore, a subcategory for atheists (non believers) was also added in this category.

The political category identified websites which promoted liberal thinking and oppose antidemocratic government ruling.

The news category identified popular international news channels (e.g. BBC, CNN, Fox).

The entertainment category identified websites having content related to music, films, gambling and such like.

Finally, the social networking category considered websites such as FaceBook and YouTube, which promote file sharing and other means of social communication. Also communication websites involving VoIP tools (e.g. Skype) have been considered here. This is because some countries control telecommunication media through government ownership and since the telecommunication media is not privatized websites which promote online communication via web technology means are regulated.

Ten websites were identified from each category, leading to a total of 100 websites which were added in the pick-list of the WBF, and tested in all of the ten countries. The tests were conducted in June 2009. The percentages of websites blocked in each identified category are shown in Table 1.

<table>
<thead>
<tr>
<th>Category</th>
<th>Percentage of Blocking</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Canada</td>
</tr>
<tr>
<td>Religion</td>
<td></td>
</tr>
<tr>
<td>Islam Shia</td>
<td>0</td>
</tr>
<tr>
<td>Islam Sunni</td>
<td>0</td>
</tr>
<tr>
<td>Christianity</td>
<td>0</td>
</tr>
<tr>
<td>Judaism</td>
<td>0</td>
</tr>
<tr>
<td>Bahai</td>
<td>0</td>
</tr>
<tr>
<td>Atheist</td>
<td>0</td>
</tr>
<tr>
<td>Political</td>
<td>0</td>
</tr>
<tr>
<td>News</td>
<td>0</td>
</tr>
<tr>
<td>Entertainment</td>
<td>0</td>
</tr>
<tr>
<td>Social networks</td>
<td>0</td>
</tr>
<tr>
<td>Avg blocking</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1. Snapshot view of website blocking

The snapshot view in Table 1 reveals that Canada, New Zealand, Turkey and US do not have blocking in the identified categories. But, varying degrees of blocking exists for other six countries. As seen in Table 1, Saudi Arabia has been found to have maximum blocking of 30%, followed by Iran at 26%, then China at 13% which is followed closely by UAE at 12%, and finally by France and Germany at 2%. However, it should be noted that these percentages do not represent absolute values as blocking may change depending upon current happenings and political situation.

Next, the study differentiates blocking existing between different regional areas by defining three regional groups – Middle East, West and East. The Middle East countries (Iran, Saudi Arabia and UAE) have been grouped together, Western countries (Canada, France, Germany, New Zealand, Turkey, and USA) in another group and the Eastern group comprises China. The question facing us was whether to consider Turkey with Middle East identity or European identity. Based upon the blocking data collected from the test run, we have considered Turkey in the Western group. We realise Turkey’s grouping raises some questions, but we felt that adding Turkey with other Middle East countries may skew our reporting. Our data showed that Turkey had no blocking in any of the five categories and had adopted similar restrictions as countries belonging to the Western group. Hence if Turkey’s data is considered with the Middle East group, this would reduce the overall blocking averages of this group drastically, but if considered with the West group would have minimal impact on their overall blocking averages. Moreover, discussions regarding Turkey’s position belonging to the Middle East or to Europe are still ongoing (Robbins, 1996).
Table 2 shows the percentage of Internet blocking found from the WBF for the three groups.

<table>
<thead>
<tr>
<th>Category</th>
<th>Middle East</th>
<th>Western</th>
<th>Eastern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Religion</td>
<td>23.9</td>
<td>1.11</td>
<td>10</td>
</tr>
<tr>
<td>Political</td>
<td>20</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>News</td>
<td>23.3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Entertainment</td>
<td>6.67</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Social Networking</td>
<td>33.3</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>Average Blocking</td>
<td>21.4</td>
<td>0.22</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 2. Website blocking based upon regional groupings

Table 2 shows that maximum blocking of 21.4% exists in Middle East countries. Our findings indicate that regulations have been applied in all categories relating to political, social, cultural and religious beliefs. The high degree of blocking of social networking sites in the Middle East countries could also be related to strict regulations imposed on telecommunication agencies through tightly controlled licensing authorities. Many governments have outlawed VoIP services due to concerns that these services (e.g., Skype) automatically encrypt voice conversations, making tapping of conversations more difficult, and also to protect existing telephony providers in the tightly controlled licensing environment (Cankorel & Aryani, 2009). The Eastern group consists of one country China, having Internet blocking of 16%. The state run agencies have blocked websites promoting political and religious views besides some social networking websites. Moreover, the regulatory practices on Google by China are currently an ongoing issue (Stone & Xin, 2010) and our study also finds that some regulations are in place. Finally, Western countries have lowest blocking of 0.22%. However, this is mainly due to website blocking in religious categories by France and Germany. The other countries in this group have no blocking whatsoever. Interestingly, Turkey, which is predominantly an Islamic country, has not been found to impose restrictions in any of the categories mentioned.

It is imperative to note that the percentages denoting Internet blocking are not constant and may change depending upon many factors. Some factors relating to Internet blocking are based upon political reporting by local and international journalists, current political situation or a country’s security concerns. Furthermore, new website domains are being created daily in the ever evolving web, further leading to a continuously changing picture of blocking rules and policies. Different innovative techniques are being discovered to masquerade IP addresses, which in turn mean that new methods are needed by government regulators to unmask websites for addition to blocked lists. Thus this dynamic game where technology is being used to fend off government imposed regulations seems to be long game which will be played for a quite some time.

7 CONCLUDING REMARKS

This study has revealed a snapshot of extent of Internet blocking in ten countries. Websites are often filtered and censored in many countries, if governments consider the website information obstructive and not aligned with their political, religious, economic or cultural sentiments. However, there are two sides of the debate. While some may consider that the Internet is self regulating and the government need not be involved, there are some others who reject this view and consider that “the government needs to be involved to ensure that the Internet does not lose the trust of users”. Internet controls are under considerable scrutiny given the “competing interests” and “complex interrelationships”, and national agencies are still of the opinion that “we don’t have the right governance tools” (Technical Report, 2010, p. 17). More recently, the Internet has become an arena for illegal downloading of files bound by copyright agreements (e.g., movies, music), and many countries are proposing filters and audits of ISP traffic to hunt down ‘Internet pirates’ (Doesburg, 2010).

Our study has shown that regulatory controls to legislate online access on the Internet are embedded in sensitive domestic issues concerning religious, cultural, social, economical and political sentiments. Religious governments enforce strict control on religious websites; cultural sentiments help in
deciding what is considered as unethical or unsocial to enforce restrictions on social networking and entertainment websites; and political views on democracy, dictatorship and socialism play a role on how to legislate websites which encourage open dialogue and discussion forums.

However, in our study, we have looked at Internet regulation from a practical viewpoint and have demonstrated that varying levels of controls are present in different country contexts. An extensive literature review has explained both sides of practice methods used to regulate Internet, that is, what practices are applied by government regulations to block websites, and what counter practices can be used to bypass regulations and un-block websites. Finally, using a software tool, we have conducted experiments to demonstrate the extent of Internet filtering in three geographical domains. A limitation of our study is a snapshot view of Internet blocking obtained at a particular point in time opposed to a longitudinal study conducted over a period of time. A longitudinal study would provide more detailed view on each country’s censorship policies. Also, the study has been restricted to ten countries in view of risks taken by participants who were accessing websites censored in their country. Whilst there could be many viewpoints on Internet regulatory practices, this study has achieved its aim in sharing some new insights on Internet blocking across ten diverse country domains.

References