RESEARCH

An Investigation of the Performance of UK Internet Providers' Web Filters

Matthew Rowe[1*] and Richard King[2]

[*]Correspondence:
m.rowe@lancaster.ac.uk
[1]Data Science Group, School of Computing and Communications, Lancaster University, LA1 4WA Lancaster, UK
Full list of author information is available at the end of the article

Abstract

Since 2013 UK-based broadband and mobile internet users have been faced with a *yes* or *no* decision in : *would they like to turn web filtering on, or not?* A minority of users (between 6 and 36%) chose to say to turn on filtering, often to protect users of their connection from potentially harmful and damaging content. Since their inception, filters provided by UK Internet Service Providers (ISPs) and Mobile Service Providers (MSPs) have been reported *anecdotally* as having both *overblocked* web sites (i.e. erroneously blocked a page when it should not have been) and *underblocked* web sites (i.e. not blocked a web page when it should have been); however until now no work has investigated the performance of UK ISPs' and MSPs' web filters and how this can be empirically characterised. In this paper we fill this gap by presenting the first systematic study of UK web filters used by consumers. By using data provided by the Open Rights Group, and performing a data science study, we were able to compute the performance of the UK ISPs' and MSPs' filters across a range of web sites, providing evidence of systematic underblocking and overblocking. Our findings are particularly stark: we observed that between 30% (for O2) and 82% (for TalkTalk) of tested URLs were underblocked, while between 2% (for EE) to 6% (for BT) of URLs were overblocked.

Keywords: Data Science; Web Filtering; Censorship

Introduction

In 2013, the United Kingdom government instructed all UK-based Internet Service Providers (ISPs) to provide new customers with an '*unavoidable*' choice: *to turn on web filtering, or not.* This was presented to such customers as a web-based form to which the customers provided their answer; and should the customer select *yes* then for certain ISPs additional questions would be asked about the level of filtering required. Such web filtering was mandated in order to protect children from adult content (e.g. pornography, alcohol, and drugs), and to ensure that they can browse and use the Web in a safe manner. However after one year of the provision of such filters, an Ofcom report[1] found that for new ISP customers, who were offered the choice of filtering uptake ranged across ISPs from only 5% to 36%.

During the period since the filters' inception, various news outlets have reported on examples of '*overblocking*' by ISPs - where sites are blocked that should not have been - such as sexual health advice blogs, charity web sites, addiction-support sites, and politics-related web sites and opinion blogs. This has led to questions being raised as to the accuracy of the filters, what they are blocking that they

[1]http://stakeholders.ofcom.org.uk/internet/internet-safety-2

should not be (*overblocking*), and what sites they are not filtering out that they should be (*underblocking*). As such, this growing discourse is calling into question the *efficacy* of the filters and the degree of censorship that they are enabling. Despite such questions being raised, at present little is known of how effective the filters are, as the ISPs do not report on their accuracy. Motivated by this current lack of understanding, in this paper we investigate the following three research questions:

1 **RQ1**: How can we understand how UK ISP Web Filters function, and how accurate they are?
2 **RQ2**: Are there certain categories and domains of web sites that are prone to overblocking?
3 **RQ3**: How long does it take an ISP to fix an error?

In order to investigate the above questions, we present a study of both UK ISPs and Mobile Service Providers' (MSPs) filters using data collected by the Open Rights Group[2] as part of their Blocked.org.uk[3] project. The aim of the project was to *probe* a range of Internet (ISPs) and Mobile Service Providers (MSPs) with a collection of URLs and collect examples of blocked and unblocked web sites. In performing this study, we follow a *data science* approach by first performing exploratory analysis at the *macro* level of what domains are commonly blocked and what categories of sites are blocked by the filters, before then investigating the accuracy of the filters and to identify any categories of sites that are routinely *overblocked* and *underblocked*, thus performing a relational study between filters and their accuracy, and site categories.

This paper makes the following contributions:

1 An empirical characterisation of UK ISPs' and MSPs' filtering accuracy over time.
2 Evidence of overblocked and underblocked sites by UK ISPs and MSPs.
3 A computational framework harnessing Apache Spark for parallelised processing of probe data.

This work is the first to investigate the performance of UK ISP web filters and to provide evidence of both overblocking and underblocking. For that reason, the work has huge potential for implications on the domains of digital rights and censorship, and also data science in the methodology that we follow in investigating web filters' performance through data - so-called '*data-driven digital accountability*'. We begin this paper by first explaining related works study web filtering technology and web censorship, and the inherent impact of both; before then moving on to outlining which Internet and Mobile Service Providers. We follow this up by describing how the *probe* system works for monitoring web filters, before then presenting evidence of what domains and categories of sites are being blocked and by whom. In the proceeding sections we then explain how we gauge filter accuracy, present qualitative examples of incorrectly blocked sites, and investigate how quickly an ISP responds to fix an incorrect block. In order to provide full transparency of how this paper's results and findings were derived, both the software used to analyse the web filters

[2] A non-profit UK-based organisation who campaign for and work to promote digital rights
[3] `https://www.blocked.org.uk/`

and the results from our analyses are available on the Open Rights Group's Github repository.[4]

Related Work

To date, the investigation and exploration of web filtering has been largely concentrated in the literature around *censorship*, despite the evocative-nature of the term. For instance, early work by Adkeniz [1] argued against censorship as rhetoric was emerging around the need to censor the Internet - largely in order to protect under-18s from being exposed to potentially harmful content. Adkeniz's view was that free speech must be maintained here, and that filtering must only be performed with the correct safeguards in place - to ensure correct application. McIntyre and Scott [2] expanded over this line of work by examining the role of web filtering and governance. The authors argued that while existing forms of censorship are mandated by politicians, web filtering follows a different route and involves different actors (e.g. third-party companies), thereby reducing the transparency surrounding the process and the accountability that accompanies this.

The rise in state censorship of the Internet in countries such as China, Syria, Saudia Arabia, and Turkey, has led to a body of work attempting to understand what filtering mechanisms are at play, and whether and how such mechanisms are being circumvented. For instance, Verkamp & Gupta [3] looked at different mechanisms by which web censorship takes place throughout different countries (e.g. Turkey, Saudi Arabia). The authors mapped out the landscape of filtering mechanisms, finding: different triggers (e.g. hostname, IP address), and modes of censorship application (e.g. filtering requests, modifying responses); in doing so, the authors were able to devise a system to probe which URLs were blocked and by whom. Similar work by Dalek et al. [5] presented a method to detect which filtering technology is being used for censorship, again focussing on state-level censorship. Their approach demonstrated that a combination of HTTP headers' keywords and path information can be used to identify known filtering technologies being used (e.g. Netsweeper).

The expanse of web filtering across states has seen the creation of community-led internet-wide initiatives to monitor censorship. One such initiative is the OpenNet Initiative (ONI)[5] run with the intention of gathering evidence of censorship and providing the technical infrastructure to monitor the use of Internet filtering. Work by Crete et al. [8] used data provided by ONI to understand how censorship is performed and why this takes place. The authors described a by-product of censorship known as '*collateral filtering*' where filtering leads to other content being blocked inadvertently - i.e. so-called overblocking. This notion of *collateral filtering* is reinforced by Murdoch & Roberts [7] when examining the role of censorship and its perceptions, as they state that: "*... over-blocking is an underhanded attempt to avoid criticism, but other times it proves to be a mistake resulting from overzealous interpretations of rules or collateral damage due to technical limitations in censorship techniques.*" Similar to ONI, the Tor project's Open Observatory of

[4] `https://github.com/openrightsgroup/cmp-analysis`

[5] `https://opennet.net/`

Network Interference (OONI) software [10] has been used throughout community-led initiatives to monitor which URLs are blocked, where and when. While Aceto et al. [6] provided a platform known as the User-based Internet Censorship Analysis (UBICA) platform to allow users to run tests over their ISP connection to ascertain what is being blocked. Given the myriad ways in which web filtering can function (DNS-tampering, keyword blocking), we refer the reader to the detailed and comprehensive review of approaches for detecting web filtering by Aceto and Pescape [9] for further information.

The study of web filtering and its mechanisms transcends various layers including the state, as in [3, 5, 8, 6], and organisations. For the latter, Esnaashari et al. [4] focussed on web filtering in New Zealand throughout organisations that provide web access- e.g. in libraries, cafes, etc.. Unsurprisingly, the authors found that different organisations blocked different types of content and applied different levels of filtering.

Computing filter accuracy

One of the core aims of our work is to understand how well UK ISPs and MSPs perform web filtering, thereby allowing the public to understand how reliable web filtering is. Prior work has sought to gauge the degree of filtering, however the limitation to state censorship restricts researchers from knowing what *should* be blocked - and thus allowing the accuracy of filters to be gauged. One of the first works in this direction was produced by researchers from Google's Zion VLab [11] who examined the extent to which collateral damage occurs through state-level censorship programmes. The authors found evidence of DNS injectors along query transit paths, meaning that routing of hostname responses is injected as a form of filtering - happening within the transit-phase of a hostname being queried and then resolved. More recent work by Nabi [12] investigated the uptake of certain web sites in countries where they have been blocked. The author demonstrated that sites that had been publicly declared as blocked actually increased in their visits post-blocking, thereby suggesting evidence of a '*Streisand effect*'.[6]

Despite the wide body of work covering the detection of filtering approaches and their usage across various countries, we could only find one piece of work that empirically characterised web filters' accuracy. The work in question, by Stark [13], used a sample of URLs categorised as either adult or clean, passed these URLs through various home PC web filters (e.g. McAfee, CyberPatrol) and then computed: (i) the underblocking percentage (i.e. the percentage of URLs that should have been blocked that weren't), and; (ii) the overblocking percentage (i.e. the percentage of URLs that were blocked that should not have been). Stark's results derived underblocking percentages ranging from 6.2% to 43.4% for various filters, and from 0.4% to 20.7% for overblocking; thereby indicating that filters perform better at minimising incorrect blocks than detecting what it should be blocked. As of writing this paper, we were unable to find any literature that examined the performance

[6] The *Streisand Effect* occurs when a given party attempts to block, or censor, information being published, and in doing so raises awareness of said information - thereby having the opposite effect of that intended.

of UK ISP and MSP web filters, nor their uptake - aside from the Ofcom Internet Safety report in 2014.[7]

The above works demonstrate that researchers have largely concentrated on understanding how state and organisation-level censorship takes place, and the myriad ways in which filtering operates at a technical level. As such, existing work has yet to quantify web filters' accuracy and the degree to which '*collateral filtering*' is evident (i.e. overblocking and underblocking); we believe that this is largely due to the lack of prescribed lists of gold standard blocks. In this paper we present for the first time evidence of such collateral filtering and provide empirical evidence of how accurate ISP and MSP web filters are. This is enabled, as we will detail below, by examining ISPs' and MSPs' descriptions of their filters and their intended categories of blocked sites, which we operationalise through modelling filters as *pseudo-classifiers* and computing their accuracy against a gold-standard.

Studied Internet and Mobile Service Providers

The United Kingdom's telecommunications market is one of the biggest in the world, and consumers are provided with a range of Internet Service Providers and Mobile Service Providers to choose from. In order to provide data for our analysis, we used both ISPs and MSPs that had the largest consumer bases in the country. For the ISPs we examined the web filtering of: BT, PlusNet, Sky, TalkTalk, and VirginMedia; while for the MSPs we examined the web filtering of: EE, O2, T-Mobile, VirginMobile, and Vodafone. We now describe the filtering technologies and available filter settings of each.

Internet Service Providers' Filters

BT. BT provide a system known as '*BT parental controls*'[8] which uses DNS-based blocking of URLs. The system utilises site categorisation information from Nominum[9] by looking up requested hostnames and blocking any requests for URLs from banned lists. BT provide three levels of filtering once turned on: (i) *light*, which blocks pornography, obscene and tasteless, hate and self-harm, drugs, alcohol and tobacco, and dating; (ii) *moderate*, which blocks all of the light filter settings plus nudity, weapons and violence, gambling, and social networking, and finally; (iii) *strict*, which blocks all of the above plus fashion and beauty, file-sharing, games, and media streaming.

PlusNet. Use network-level filtering of piracy sites, and at present do not filter any adult web content.

Sky. Sky provide a filtering system known as '*Sky Broadband Shield*' that also uses DNS-based blocking of URLs, however unlike BT, Sky's system uses site categorisation information provided Symantec and their Rulespace Web Content categorisation system.[10] Similar to BT, Sky offer three levels of categorisation: (i) *18* which

[7] http://stakeholders.ofcom.org.uk/internet/internet-safety-2

[8] http://www.productsandservices.bt.com/products/manage-broadband-extras/

[9] http://nominum.com/

[10] http://www.symantec.com/page.jsp?id=rulespace

blocks malware sites; (ii) *13* which blocks cyber-bullying, pornography, suicide and self-harm, drugs, dating, and malware sites, and; (iii) *PG* which blocks all of the above plus social networking and online gaming.

Talk Talk. Talk Talk provide a filtering system known as HomeSafe[11], however unlike BT and Sky, this system using Deep-Packet Inspection to examine URLs being visited by users. The filter includes the setting of a *Kids Safe* filter that allows certain categories of sites to be blocked: Dating, Drugs, Alcohol and Tobacco, File Sharing Sites, Gambling, Games, Pornography, Social Networking, Suicide and Self-Harm, and Weapons and Violence.

VirginMedia. This ISP provides a a system known as '*Web Safe*'[12] that is a DNS-based system which matches requested URLs with known blocked URLs in a DNS-lookup table. As with BT, VirginMedia also use site categorisation information from Nominum. From VirginMedia's web site, it is not clear what *Web Safe* blocks, therefore such categories of sites were obtainted from the OFCOM Internet Safety Measures report from 2014.[13]

Mobile Service Providers' Filters

Unlike Internet Service Providers, Mobile Service Providers can sell their products (e.g. pay as you go phones) to people under the age of 18. As a result, customers are able to access content that may not be deemed suitable for their age group. Below we describe the filtering approaches taken by the MSPs and what they block. In general, MSPs have filtering turned on by *default* and require customers to turn off the filters by verifying their age.

EE and T-Mobile. Both EE and T-Mobile use a system known as '*Content Lock*'[14]. This system has one filter setting that blocks categories of sites including: alcohol, anonymisers, criminal skills, drugs, gore, hacking, hate, pornography, self harm, sex advice, suicide, tobacco, and violence.

O2. Use a filtering system known as '*O2 18+*' with default-on setting, so filtering is always applied. Symantec RuleSpace[15] is used for site categorisation, and anything that is classed as '*Adult*' by the BBFC (British Board of Film Classification)[16] is blocked.

Three. As with O2, blocks any sites categorised as Adult, and therefore suitable for those over the age of 18, by the BBFC.

---

[11] http://www.talktalk.co.uk/security/homesafe-demo.html

[12] http://my.virginmedia.com/my-apps/websafe.html

[13] http://stakeholders.ofcom.org.uk/binaries/internet/internet_safety_measures_2.pdf

[14] http://ee.co.uk/help/safety-and-security/my-digital-life/content-lock-and-orange-safeguard

[15] http://www.symantec.com/page.jsp?id=rulespace

[16] http://www.bbfc.co.uk/

Vodafone. Vodafone provider filtering through their '*Content Control*'[17] system, with categorisation of web sites provided by Symantec. Filtering is default-on, meaning that customers must remove filtering manually by proving that they are over 18. There is only one level of filtering used here and this blocks: chat and dating services, erotica, gambling, and violent games.

Monitoring Web Filters

We now move onto the crux of our work. In order to understand how *well* ISPs' and MSPs' web filters perform we first need to gather information about which URLs each filter has blocked and when this took place. For this we used data provided by the Open Rights Group as part of their Blocked.org.uk project: below, we describe hos the project's probe system functions before then moving on to describing the data that we collected.

Blocked.org.uk Probe System

-Explain the framework that was used for this -Explain the submission interface and the use of the blocked portal to check what has been blocked and unblocked -Gathering evidence of overblocking and underblocking -Explain the settings of each filter - i.e. default settings for each

Data Captured

For this paper we used data collected from the probe system up to the end of March 2015, thereby providing around 9 months' worth of data for analysis. Figure 1 shows the distribution of requests per day that were collected in our dataset - a single request consists of a URL being tested by a given filter at a given time, hence representing a unique record. In total we collected 6,289,550 requests for URLs across the studied ISPs and MSPs; Table 1 shows the requests per-ISP and per-MSP that were collected.

Table 1 Dataset of requests from Blocked.org.uk

	Time Interval	Requests
BT	[July 2014, April 2015]	877,768
PlusNet	[June 2014, April 2015]	876,218
Sky	[June 2014, April 2015]	883,571
TalkTalk	[July 2014, April 2015]	883,787
VirginMedia	[June 2014, April 2015]	883,743
EE	[May 2014, March 2015]	270,471
O2	[May 2014, March 2015]	495,429
Three	[May 2014, March 2015]	282,017
T-Mobile	[April 2014, September 2014]	483
Vodafone	[March 2014, March 2015]	836,063

Inspecting the per-ISP distributions of requests per day, as shown in Figure 2, we can see that we cover different time intervals for different ISPs with BT's and TalkTalk's collection starting later than PlusNet, Sky and VirginMedia. Such differences in intervals were due to the slight delay in configuring the probe system for these connections.

[17] http://support.vodafone.co.uk/Internet/Content-control-and-Vodafone-Guardian/38914008/How-can-I-switch-off-content-control-so-I-can-access-age-restricted-content-on-the-internet.htm

Figure 1 Number of URL requests made over time since the beginning of the project.

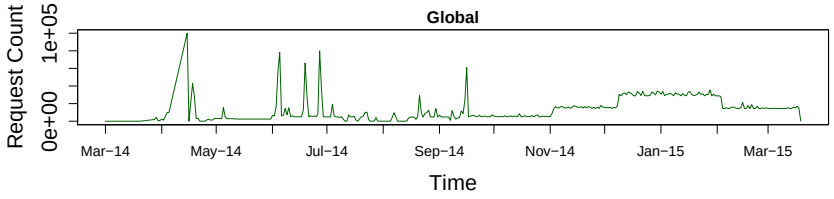
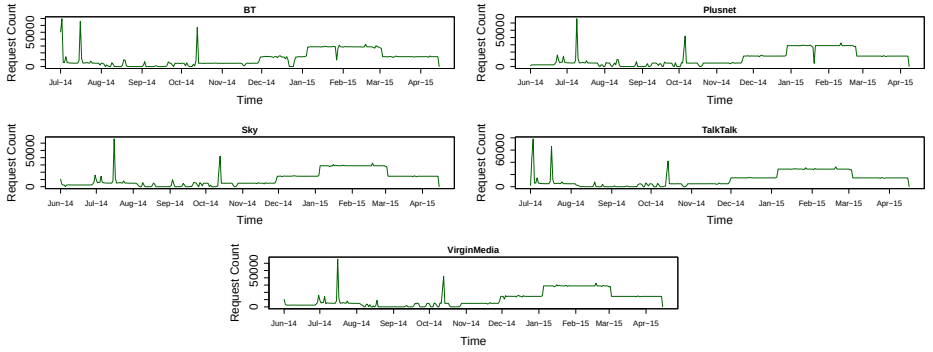
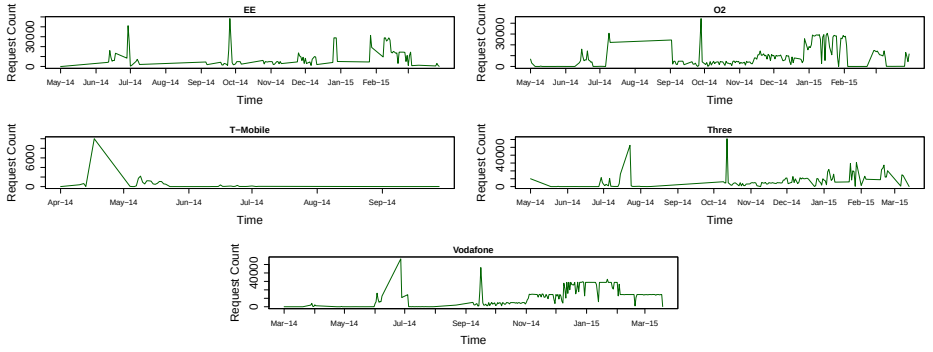


Figure 2 Number of URL requests made per broadband ISP filter



In a similar vein to the ISPs' plots, we can also inspect the distribution of requests per day for the MSPs - shown in Figure 3. As with the ISPs, we observe that collections cover different time intervals: Vodafone collection spanning March 2014 to March 2015, EE, O2 and Three spanning May 2014 to March 2015, and T-Mobile spanning April 2014 to October 2014. This latter MSP's probe had issues collecting requests and testing for blocks, therefore we do not use T-Mobile in our analysis - despite this we are able to analyse EE which uses the same underlying filtering technology.

Figure 3 Number of URL requests made per mobile ISP filter



Blocked Content

This section begins our exploratory analysis of what has been blocked by the UK ISPs and MSPs. We start by assessing the top-level URLs domains of blocked requests, before then moving on to examine the categories of blocked URLs.

Blocked Domains

We began by investigating which domains had been blocked by Internet Service Providers. To do this, we extracted the blocked requests for each ISP and then recorded the frequency distributions of the blocked URLs domains (e.g. `www.lancs.ac.uk/staff/rowem` would have domain `lancs.ac.uk`).

Figure 4 shows the top-40 domains that are blocked by ISPs' filters. As PlusNet does not use filtering, we can see that the only URLs that are blocked are those related to file-sharing sites such as The Pirate Bay and similar platforms - these were served with a UK court order to be blocked in 2013. For the remaining ISPs, the top-most domains that are blocked are pornography (e.g. `pornhub`, `xhamster`, etc.), and other adult-content sites (e.g. `4chan`). Interestingly, we found that certain domains are picked up that we would not associate with adult content, such as `reddit`, `tumblr`, and (worryingly) `torproject`.

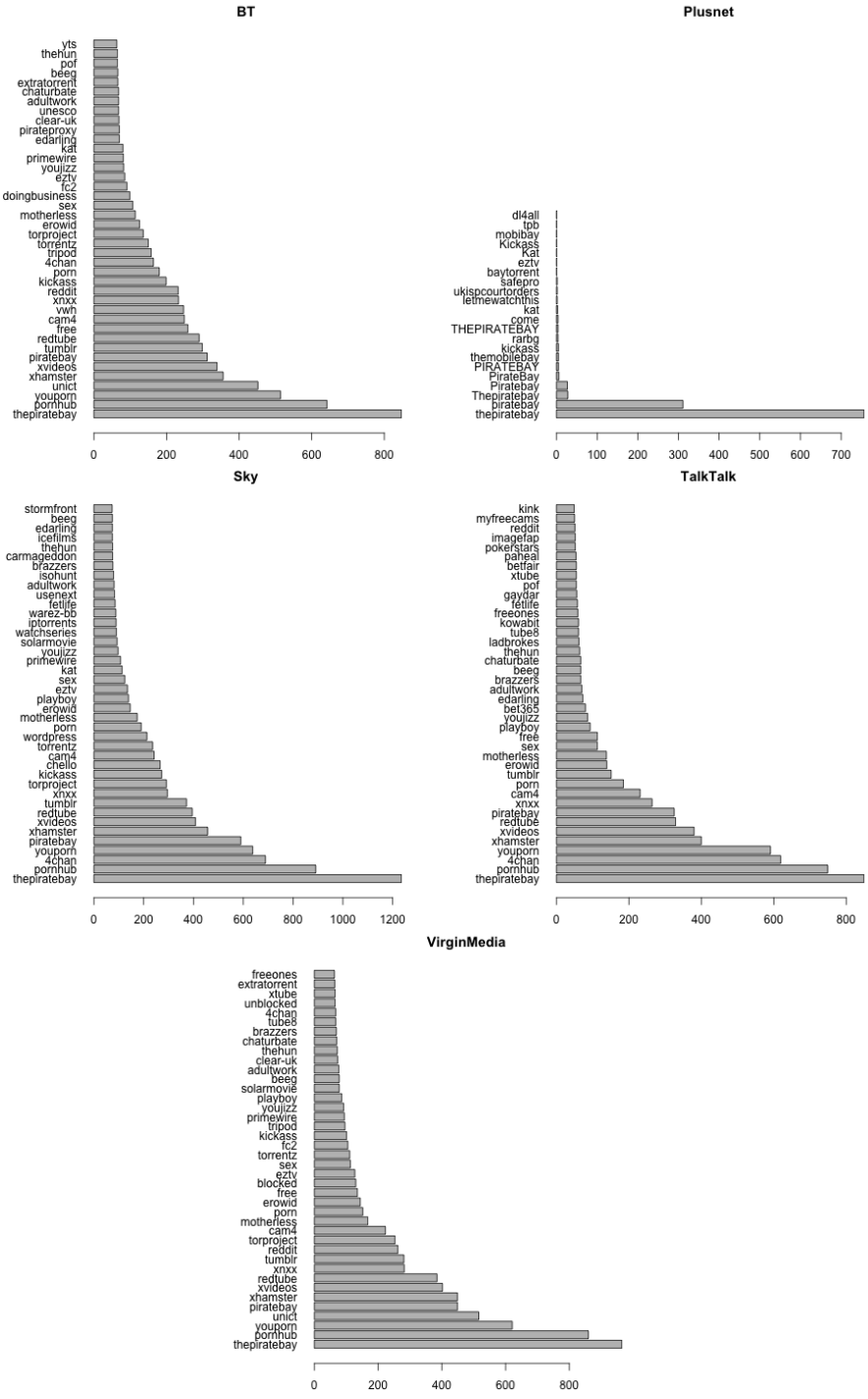
Inspection of the top-40 blocked domains of blocked sites by MSPs (in Figure 5) reveals similar findings to that of the ISPs: with pornography and adult-content sites being blocked, yet with other domains of sites that are not normally associated with adult-content also being blocked (e.g. `torproject`, `tripod`, `wordpress`).

Overblocked Sites

Our macro-level analysis of blocked domains revealed that certain filters were blocked site domains that are not normally associated with adult-content. To examine further what these sites were, and to find out if they indeed contained adult content, we extracted blocked sites with domains from `wordpress`, `tumblr`, `reddit`, and `livejournal`; we now describe *overblocked* sites from each (i.e. sites that should not have been blocked by the filters):

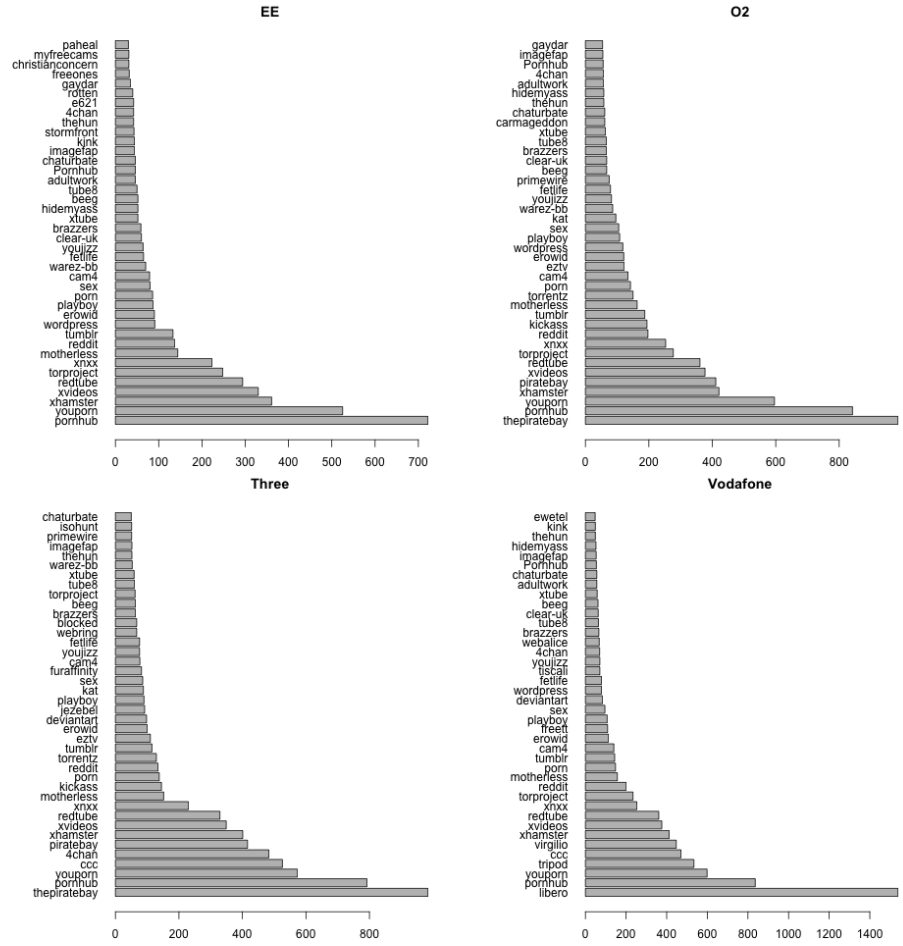
- `http://garsai.wordpress.com` - Radio documentary site. Blocked by TalkTalk.
- `http://toysoldier.wordpress.com` - This site is a support site for men who have been abused. Blocked by Sky and O2.
- `http://www.heyartist.wordpress.com` - Site promoting art as a support mechanism for enhancing wellbeing. Blocked by Sky.
- `http://azurelunatic.tumblr.com/post/18654147576/ive-been-forced-to-explain-homosexuality-` - Page explaining why someone is gay. Blocked by Vodafone, O2, TalkTalk, and BT.
- `http://thusly.tumblr.com` - Personal web blog containing artistic materials and shared music. Blocked by EE, BT, Sky, O2, and Vodafone.
- `http://reddit.com/r/creepypms` - Subreddit sharing creepy private messages that people have received. Not necessarily adult content, and definitely not pornography. Blocked by EE.
- `http://community.livejournal.com/asi/` - Anorexia and self-harm support community site. Contains posts from people explaining their afflictions and getting support from other people. Blocked by Sky.

Figure 4 Distributions of blocked domains for broadband ISPs This presents which domains each ISP has blocked and the frequency distribution of those domains.



- `http://urban-decay.livejournal.com` - Photos of urban areas that have fallen into decline. Blocked by Sky, Three, and O2.
- `http://ercasse-ainince.livejournal.com/30230.html` - Article about films that have been out for a long time. Blocked by Sky.

Figure 5 Distributions of blocked domains for mobile ISPs This presents which domains each ISP has blocked and the frequency distribution of those domains.



Blocked Site Categories

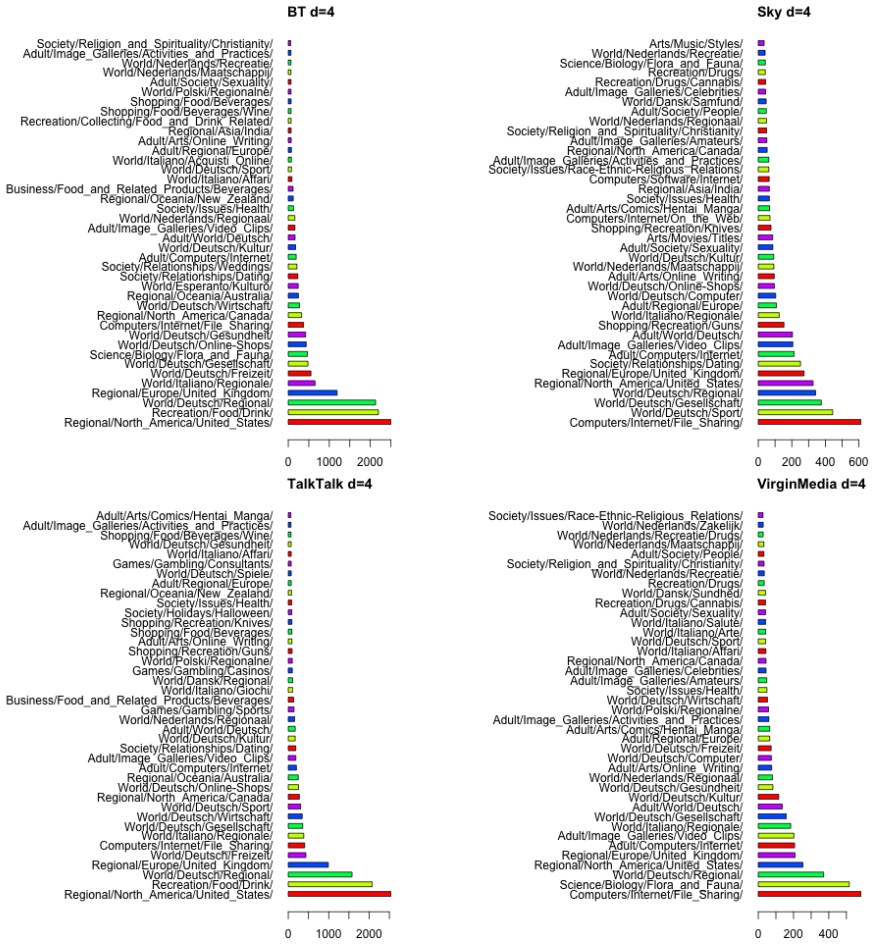
Examining the domains of blocked sites allows for an insight into whether sites are being *overblocked*, while our inspection of specific domains' URLs revealed that certain sites and pages were indeed being blocked when they should not have been - based on the ISPs' and MSPs' filter descriptions. However, manual inspection of the sites from the blocked records, in to understand what content they contain, is not feasible given their large scale (33,297). Therefore to overcome this issue, we harnessed DMOZ (Directory MOZilla)[18] which is based on the original Open Directory Project. DMOZ provides a community-created and maintained categorisation system of web sites: users submit URLs of sites and tag the category (from a previously defined hierarchy) that the sites belong to (one category per site). The submissions are then verified by the community, thereby ensuring that sites are labelled appropriately.

The DMOZ categorisation system uses ODP categories which can extend to several levels within the category taxonomy's hierarchy, therefore we were able to

[18] `http://www.dmoz.org/`

examine the distribution of blocked categories of sites up to a given depth d. This process involved looking up, for each ISP and MSP, the blocked sites' categories and then recording the frequency distribution of the categories. Below we show the top-40 blocked categories up to a depth of 4 for ISPs (excluding PlusNet from our analysis as this ISP only blocks file sharing platforms) - in Figure 6.[19] As expected, several adult categories are found within the top-40 covering different subtrees within the DMOZ taxonomy; however categories of sites appear to the blocked that one would not associate with adult-content (as with the domains above), such as: `Recreation/Food/Drink`, `Art/Music/Styles`, `World/Deutsch/Computers`, etc. However, as we will show below, several sites are placed within these categories which are, indeed, adult in nature and should be blocked as per the filter settings - i.e. covering topics such as alcohol.

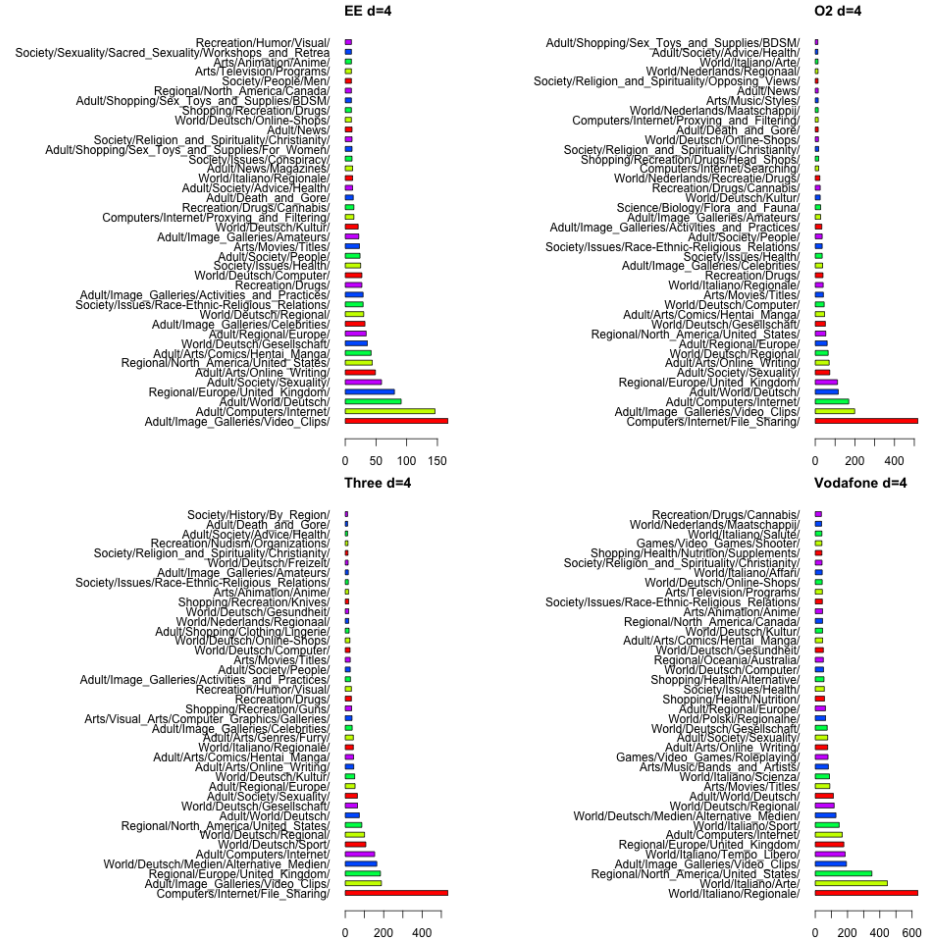
Figure 6 Distributions of blocked level-4 categories for broadband ISPs.



[19] N.b. We used a depth of 4 here in order to show how specific the categories can be, while not being too specific. We have also produced the blocked categories for depths of 3, 5 and 6, the plots of which can be found in the github repository - `https://github.com/openrightsgroup/cmp-analysis/tree/master/plotting-scripts/plots/per-isp-filter`

We found similar distributions of blocked categories for the MSPs as we did for the ISPs - as shown in Figure 7.

Figure 7 Distributions of blocked level-4 categories for mobile ISPs.



Gauging Filter Accuracy

Thus far our inspection of ISPs' and MSPs' filters has been through exploring the domains and categories of blocked sites, and has revealed *qualitative* examples of *overblocking*. We now turn to answering our first research question: *How can we understand how UK ISP Web Filters function, and how accurate they are?* To do this we proposed an approach that constructs *Pseudo Classifiers* for each ISP and MSP filter. These psuedo classifiers are then used to identify what *should* be blocked and what *should not* be blocked, based on the filters' settings' descriptions; in doing so, we can then gauge the *accuracy* of each filter. We begin this section by explaining how we construct the pseudo classifiers for each filter.

Pseudo-Classifiers for ISP and MSP Filters

Construction of the pseudo-classifier for each ISP and MSP used the DMOZ categorisation system in conjunction with each filter's documentation to understand

what should be blocked and by whom. In order to aid comprehension of what should be blocked Table 2 collates each filter's settings together with the categories of blocked sites. By default, pornography and obscene and tasteless content (e.g. gore, death, etc.) is blocked across all filters, ISPs all block hate and self-harm sites while MSPs do not, and ISPs vary in which categories of content they block - with TalkTalk being the most strict of all the filters.

Table 2 Categories that are blocked by each ISP and MSPs' pseudo-classifier according to filter documentation.

	BT	Sky	TalkTalk	VirginMedia	EE	O2	Three	Vodafone
Pornography	×	×	×	×	×	×	×	×
Obscene and Tasteless	×	×	×	×	×	×	×	×
Hate and Self-harm	×	×	×	×				
Drugs	×	×	×	×	×			
Alcohol	×		×		×			
Tobacco	×		×		×			
Dating	×	×	×					×
Social Networking		×	×					
Gambling		×	×					×
File-sharing			×	×				
Violence			×					
Gaming			×					
Malware		×	×					

In order to operationalise these categories, we identified equivalent DMOZ categories that sites had been mapped to. For the DMOZ categories we selected the most *general* category possible, therefore should a site be placed in a sub-category then this would be detected. For instance, for the category *Tobacco* we used the DMOZ categories `Shopping/Tobacco` and `Recreation/Tobacco`. Our approach was to be as conservative as possible here and to ensure that a filter would block any category that we were unsure about - e.g. blocking all sites listed under `Computers/Hacking`. Certain categories of sites were ambiguous and could contain content that should be blocked and content that should not be blocked. For instance, we found that several alcohol and tobacco related web sites were contained in the `World` category, therefore we excluded this category of site completely from our analysis. Likewise, we also excluded any sites categorised under `Computers/Software/Internet/File_Sharing` as this was also found to be ambiguous. For a full listing of the blocked categories and their DMOZ category mappings please refer to the appendices.

Judging Filter Accuracy

Once the pseudo-classifiers were constructed for each filter we could then judge how well each filter performs *overall* before then assessing how accuracy changes over time. To judge the accuracy of the filters we borrow five measures of performance from machine learning and supervised classification tasks in particular. In such classification, the goal is to induce a model that can, as accurately as possible, differentiate between classes of objects (e.g. types of customers, predicting rain or not the following day, etc.). In the context of our analyses we have a similar task: *should a given web site be blocked or not?* Therefore, we can resolve each filter to handle a set of web sites where we compare each filter's actual labels for the sites (blocked or not) against whether the sites should have been blocked or not. This can be summarised in the following contingency table:

In using the above formulation we can count how many False Positives as the number of URLs that were incorrectly blocked; and False Negatives as the number

Table 3 Contingency table for deriving filters' accuracy measures.

		Gold Standard	
		Block	Not block
Outcome	Block	True Positive (TP)	False Positive (FP)
	Not block	False Positive (FP)	True Negative (TN)

of URLs that were incorrectly not blocked. Hence, the former measures the extent to which *overblocking* occurs while the latter gauges the magnitude of *underblocking*. In order to gauge the relative performance of the filters, we use the following accuracy measures: (i) Precision, gauges the proportion of URLs labelled as *blocks* that were correct; (ii) Recall, to measure the proportion of blocked URLs that should have been detected; and (iii) F-measure (F1) as the harmonic mean between precision and recall. These accuracy measures are defined explicitly, using the contingency table defined in Table 3, as follows:

$$precision = \frac{|TP|}{|TP| + |FP|}, \quad recall = \frac{|TP|}{|TP| + |FN|}, \quad F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{1}$$

We also define two additional accuracy measures as follows: the first of the False Positive Rate (FPR) which measures the proportion of false positives that are produced by the filter, defined as:

$$FPR = \frac{|FP|}{|FP| + |TN|} \tag{2}$$

Hence, if the magnitude of TN increases and FP decreases then the classifier produces fewer overblocks. For precision, recall, F1 and FPR all values are within the closed interval $[0, 1]$ with a value of 1 indicating perfect performance for precision, recall, and F1 - and the opposite for FPR (as we wish to minimise this value).

We also included the *Matthews Correlation Coefficient* as our fifth and final accuracy measure. This measure produces a value in the closed interval $[-1, 1]$ with: 0 indicating that the filter matches a random guesser's accuracy, 1 indicating perfect performance far superior to a random guesser, and a value of -1 indicating that all outcomes of the filter were incorrect.

Parallelised Computation

As we were dealing with 8 different filters, and their pseudo-classifiers, and had to process 6 million records - which are growing in size daily as the probes collect more data - we constructed a parallel-processing framework to compute the accuracy measures for each filter. This framework was written using the Apache Spark stack and utilised HDFS from the Apache Hadoop stack - the Python code can be found online in the Github repository of the project.[20] Computation follows a pipeline approach as follows:

1 Clean and load data into HDFS. The blocked.org.uk export file containing all requests is cleaned to reduce duplicate records and then loaded into HDFS. The DMOZ category files are also loaded into HDFS.

[20] `https://github.com/openrightsgroup/cmp-analysis`

2 Load DMOZ category map into cluster memory. The DMOZ category file is processed using a map-reduce job thereby producing `(key, value)` pairs where the keys are URLs and the values are the DMOZ category - one URL has one category. We actually load the non-adult category map and the adult category map separately - as DMOZ provides different files for each - and then perform `union` operation to join the maps. This joined map is then broadcast to the cluster so that all nodes can access the map from memory.

3 Compute contingency tables per ISP and MSP. A second map-reduce job is run to produce (key, value) pairs from the cleaned request export file in HDFS. Keys here are the ISP and MSP names (e.g. BT, O2, etc.) and values are contingency table objects. The *map*-stage functions by reading in one line from the export file and then checking if the request is a block or not and for which ISP it pertains to, a new contingency table is then produced for the (key, value) pair. The *reduce*-stage then combines all of the contingency tables together for the given filter.

4 Accuracy computation. After the final map-reduce job is complete, the above accuracy measures are then computed for each ISP and MSP based on the combined contingency tables.

The above approach is an effective way of quickly processing the requests file and gauging filter accuracy. As we will explain below, this framework also allows for time-series computation of filter accuracy to be performed.

General Accuracy

From the computational framework, we were able to compute the accuracy of each filter as shown in Table 4, where we find large differences in the accuracy levels of the filters. For instance, several filters achieve recall exceeding 0.5 indicating that they are able to detect around half of pages that *should* have been blocked, while precision remains relatively low across all filters. This latter metric indicates that a large proportion of pages are blocked when they should not be; hence, *overblocking* is taking place across all filters. There are numerous examples of overblocked sites, however to provide a single qualitative example to explain the issues that filters have we can inspect the treatment of the web site to help lesbian, gay, bisexual and transgender people stop smoking `http://www.lgbtquitsmoking.com`. This site is blocked by BT's filter as it describes tobacco and smoking, however the purpose of the site is not to promote smoking but instead to improve the health of people who smoke by quitting. Despite the extent of overblocking and underblocking, we found that the Matthews Correlation Coefficients were all in excess of 0 thereby indicating that all filters perform better than random guessing. We have uploaded the collection of sites which are false positives and false negatives to the Github repository for all of the above filters.[21]

Earlier in the related work section we explained how the prior work of Stark [13] found underblocking percentages ranging from 6.2% to 43.4% for various filters, and from 0.4% to 20.7% for overblocking. The former measure is comparable to the False Negative Rate (FNR), which is computed as $1 - recall$, which we find to range between 30% (for O2) and 82% (for TalkTalk); while the latter measure is

[21] `https://github.com/openrightsgroup/cmp-analysis/tree/master/data/output`

Table 4 Accuracy levels of the various ISPs' and MSPs' filters across the five measures.

	Precision	Recall	FPR	MCC	F1
BT	0.418	0.619	0.006	0.505	0.499
Sky	0.191	0.236	0.003	0.210	0.211
TalkTalk	0.483	0.183	0.005	0.287	0.266
VirginMedia	0.112	0.512	0.002	0.239	0.184
EE	0.281	0.637	0.002	0.422	0.390
O2	0.218	0.699	0.002	0.390	0.333
Three	0.184	0.633	0.004	0.340	0.285
Vodafone	0.083	0.568	0.003	0.216	0.144

comparable to our FPR measure (as we gauge the proportion of pages that should not have been blocked that were), where we find an *overblocking* percentage ranging from 2% (for EE) to 6% (for BT). These results paint a stark picture of the accuracy of the web filters: we find here that underblocking happens to a large degree so that content which should be filtered out is instead missed, while overblocking is generating *collateral filtering* of up to 6% of web content.

Accuracy over Time

Our *general* inspection of how well the web filters perform was carried out over the entire dataset. One question that came to mind when conducting this analysis was whether the filters *improved* in accuracy over time. To investigate this we repeated the above per-ISP and per-MSP computation of accuracy measures, however this was instead computed on a *weekly* basis. This was performed by computing *time-specific* contingency tables and then computing the above measures from the tables. We again used Apache Spark for this and wrote a map-reduce job that produced (key, value) pairs where the keys were the ISPs and MSPs' names and the values were also (key, value) pairs - where the keys were this time week numbers and the values the week-specific contingency table.

Figure 8 and figure 9 show the accuracy of the ISPs' filters and MSPs' filters over time, respectively - these values have been smoothed using a moving-average model of order 5. We find that, in general, model accuracy degrades over time, in particular the ISPs. The reason for this degradation is likely due to web site owners in the UK submitting URLs to be tested to the probe-system's queue, as the blocked.org.uk site gained more attention around 6 months after its launch. Initially, sites from the Alexa-ranking system were used for testing however as more bespoke, less mainstream, sites were added the filters were unable to categorise them effectively.

Figure 8 Accuracy of Broadband ISPs' filters over time. ARIMA(0,0,5) plot of the accuracy measures: precision, recall, f-measure (F1), false positive rate (FPR), and the Matthews' Correlation Coefficient. Weeks are from the start of the ISP-specific filter logging period.

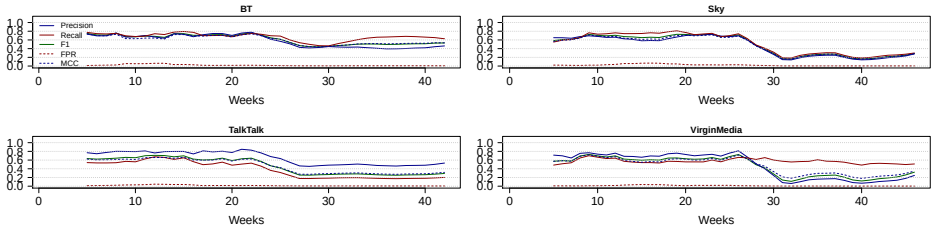
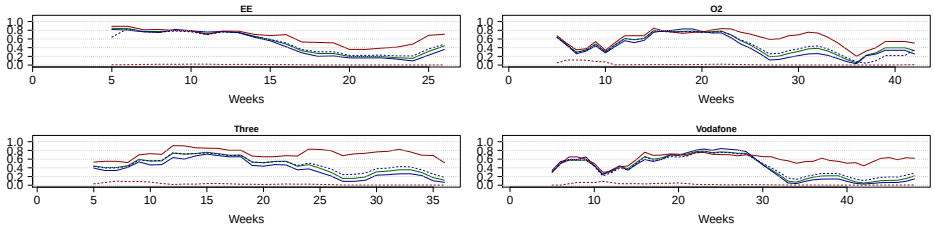


Figure 9 Accuracy of Mobile ISPs' filters over time. ARIMA(0,0,5) plot of the accuracy measures: precision, recall, f-measure (F1), false positive rate (FPR), and the Matthews' Correlation Coefficient. Weeks are from the start of the ISP-specific filter logging period.

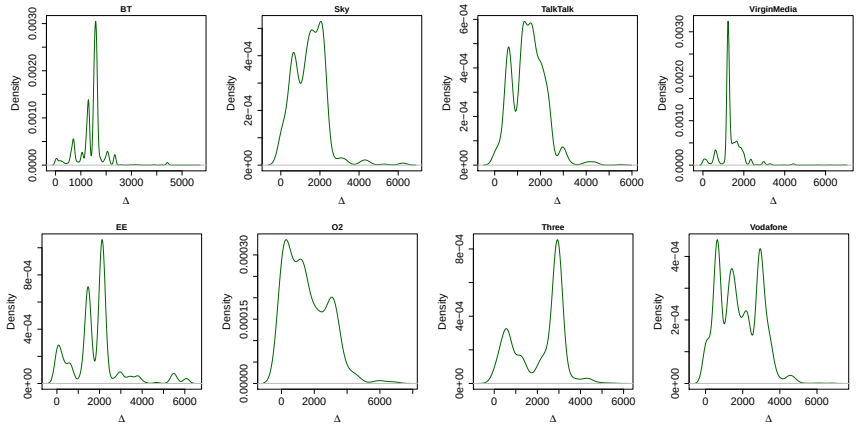


Time to Correction

Once a site has been blocked, and providing that the site owners is aware of such a block, several ISPs provide functionality to report such an error, with the view that the site would then be *unblocked*. To investigate our third research question *How long does it take an ISP to fix an error?* we examined the Δ-distribution of URLs transitioning from a *block* to an *ok* state. The distribution was computed by recording the frequency distribution of the number of days each ISP's block-to-ok transitions. Of course, one of the limitations of the probe system is that it cannot check each URL for a block every day, instead each URL is tested across the probes every 45 days (or once every 1080 hours).

Figure 10 presents the Δ-distributions for the 8 filters. We can see how these distributions all have varying multimodal shapes, largely due to the scheduling of the probe system. That said, all of the distributions' densities are concentrated above 1080 hours, the average number of hours of a given URL to be tested again. This suggests a delay in fixing the blocking errors, in particular for Vodafone which has a peak at 3600 hours (145 days).

Figure 10 Time to unblock (Δ) distribution of each Broadband and Mobile ISP in hours.



Study Limitations

In this paper we have presented a study of UK Internet Server Providers' and Mobile Service Providers' filters and how well they perform. We followed a data science approach here by using data provided by the Open Rights Group from their `blocked.org.uk` project to first explore general domains and categories of sites and to find to qualitative examples of overblocking, before then investigating how well the filters perform through the creation of pesudo classifiers for each filter. In this section we reflect on potential limitations of our study and what effects these may have had.

Measurement of Blocks

The requests export from the probe system contains equivalent URLs listed under different states at different times: i.e. a URL can be defined as *blocked* at one point in time and then *ok* later, once it has been unblocked. This issue is hard to counteract

one must either decide to not count the URL at all, or count it in both sets when computing the contingency table. We opted for the latter, therefore place one URL in a respective set of the contingency table (e.g. false positives, false negatives, true positives, true negatives), and as we use sets a URL is only counted once within the respective set.

Another potential limitation of our study and data used is the probe system itself. As the accuracy over time and time-to-unblock plots have shown, the restriction on the queueing system to only request a URL every 45 days means that analysing the changes in filters over time is somewhat limited. Despite this, we can still observe large-time spans between a URL first being blocked and then transitioning to an *ok* state.

Limitations of Pseudo-Classifiers

The use of pseudo-classifiers relies on URLs having been mapped to DMOZ categories, however this is not true for all URLs. In fact, we have coverage of 85% indicating that we do not consider 85% of URLs in our analysis. This could potentially impact our accuracy measures towards favouring known sites for which DMOZ categories exist, one could argue that the measures are likely to be an *upper bound* on accuracy and that the true accuracy is potentially lower than what we have observed.

Another issue that we encountered when using the DMOZ categorisation scheme was the prevalence of *ambiguous* categories. For instance, we repeatedly found that non-UK sites for wineries and breweries were categorised under the `World` category, despite them being about alcohol. We counteracted this by removing the `World` category completely from our analysis, however it is likely that other ambiguous categories exist and could impact our results.

Conclusions

This paper has been concerned with studying the accuracy of UK ISP and MSP filters. To focus our investigated we studied three research, which we now reflect upon.

RQ1: How can we understand how UK ISP Web Filters function, and how accurate they are?

We have presented a computational framework that constructs pseudo-classifiers for each ISPs' and MSPs' filters with the DMOZ categories of web sites that *should* be blocked. By using data provided by the Open Rights Group as part of their `blocked.org.uk` project we were able to compare the outcomes of web filters with what should have been blocked through these pseudo-classifiiers. Through constructing *contingency tables* we were able to show the large extent to which both overblocking and underblocking occurs across the filters.

RQ2: Are there certain categories and domains of web sites that are prone to overblocking?

Through inspecting the provided data we found qualitative examples of domains that are prone to overblocking (e.g. LiveJournal). Upon engineering the pseudo-classifiers, we found wide variations between the categories of sites that the different ISPs and MSPs block, therefore our future work will be focussing on the former

part of the above question in order to further investigate category-specific accuracy levels.

RQ3: How long does it take an ISP to fix an error?

We presented an inspection of the Δ-distributions of each ISPs' and MSPs' filters in which we computed the probability density function of the number of hours it takes for a site to transition from a *block* state to an *ok* state. The probe system re-tests each URL every 45 days (1080 hours), however we found the distributions' densities to be largely concentrated above this value, thereby indicating that URLs take several re-tests before they are found to have been fixed.

Implications

In this paper we have presented the first work to examine how well UK ISPs' and MSPs' web filters perform when blocking content. The results show *empirically* that the filters are error prone and that there is systematic evidence of both *overblocking* and *underblocking*: between 30% (for O2) and 82% (for TalkTalk) of tested URLs are underblocked, while between 2% (for EE) to 6% (for BT) of URLs are overblocked. These findings have implications not only on other researchers seeking to investigate how well filters perform, but also on the UK public in general. The UK media has repeatedly reported on erroneous blocking, however this growing body of evidence was largely anecdotal in nature; our work is the first to systematically critique the filters, and by following a data science methodology we are able to harness data provided by the Open Rights Group to perform this study. In order to ensure full transparency of our work, we have provided the Python and R source code used to run all of the above experiments and produce the above plots on Github;[22] we have also uploaded the evidence of overblocking and underblocking from our experiments to demonstrate what we have found.[23]

Future Work

Future work is planned in the following three areas: Firstly, we are currently working on a manual annotation system for members of the public to label URLs with the appropriate category labels. Our goal here to have at least three raters per URL and then use their decisions to decide on what the categories the URL should be under. The goal here is to have an alternative gold standard from which to derive the accuracy of the filters from - thereby providing an alternative to the existing DMOZ category system.

The second area of future work will investigate text mining as a means to identify the topics of web pages. By using the manually annotated URLs from above, these will be used to examine how well we can identify the categories of pages automatically; thereby producing an automated categorisation system that can be used to scale up the labelling, if sufficiently accurate.

The third and final area of future work is a planned follow-up study in one year's time. This study will repeat the approach presented within this paper and compare the findings, thereby informing how well the filters have performed since then. As all

[22] https://github.com/openrightsgroup/cmp-analysis
[23] https://github.com/openrightsgroup/cmp-analysis/tree/master/data/output

our code is open source and engineered to perform parallel computation, repetition of the study is not expected to be challenging.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
Open Rights Group for providing the data and engineering the solution

Author details
[1]Data Science Group, School of Computing and Communications, Lancaster University, LA1 4WA Lancaster, UK.
[2]Open Rights Group, Free Word Centre, 60 Farringdon Road, EC1R 3GA London, UK.

References

1. Akdeniz, Y.: Internet content regulation: Uk government and the control of internet content. Computer Law & Security Review **17**(5), 303–317 (2001)
2. McIntyre, T.J., Scott, C.: Internet filtering: rhetoric, legitimacy, accountability and responsibility (2008)
3. Verkamp, J.-P., Gupta, M.: Inferring mechanics of web censorship around the world. Free and Open Communications on the Internet, Bellevue, WA, USA (2012)
4. Esnaashari, S., Welch, I., Chawner, B.: Restrictions affecting new zealanders' access to the internet: A local study. In: Advanced Information Networking and Applications (AINA), 2014 IEEE 28th International Conference On, pp. 771–774 (2014). IEEE
5. Dalek, J., Haselton, B., Noman, H., Senft, A., Crete-Nishihata, M., Gill, P., Deibert, R.J.: A method for identifying and confirming the use of url filtering products for censorship. In: Proceedings of the 2013 Conference on Internet Measurement Conference, pp. 23–30 (2013). ACM
6. Aceto, G., Botta, A., Pescapè, A., Feamster, N., Awan, M.F., Ahmad, T., Qaisar, S.: Monitoring internet censorship with ubica. In: Traffic Monitoring and Analysis, pp. 143–157. Springer, ??? (2015)
7. Murdoch, S.J., Roberts, H.: Internet censorship and control [guest editors' introduction]. Internet Computing, IEEE **17**(3), 6–9 (2013)
8. Crete-Nishihata, M., Deibert, R., Senft, A.: Not by technical means alone: the multidisciplinary challenge of studying information controls. Internet Computing, IEEE **17**(3), 34–41 (2013)
9. Aceto, G., Pescapé, A.: Internet censorship detection: A survey. Computer Networks (2015)
10. The Tor Project. Ooni: Open Observatory of Network Interference. `https://ooni.torproject.org/` Accessed 25/6/2015
11. Anonymous: The collateral damage of internet censorship by dns injection. SIGCOMM Comput. Commun. Rev. **42**(3), 21–27 (2012). doi:10.1145/2317307.2317311
12. Nabi, Z.: Censorship is futile. arXiv preprint arXiv:1411.0225 (2014)
13. Stark, P.B.: The effectiveness of internet content filters. University of California, Berkeley (2007)

Appendix 1: DMOZ Block Categories
In order to operationalise the blocked categories we identified equivalent DMOZ categories. This allowed each filter to have a pseudo-classifier constructed which contained the categories of sites that should be blocked upon request. These DMOZ categories were as follows:

- Adult and Obscene and Tasteless: `Adult`
- Hate and Self-harm: No category
- Drugs: `Recreation/Drugs`
- Alcohol: `Recreation/Food/Drink/Drinking, Recreation/Food/Drink/Mead, Recreation/Food/Drink/Wine, Recreation/Food/Drink/Beer, Recreation/Food/Drink/Alcopops, Recreation/Food/Drink/Cider, Recreation/Food/Drink/Cocktails, Recreation/Food/Drink/Liquor, Recreation/Food/Drink/Sake, Health/Specific Substances/Alcoholic Beverages`
- Tobacco: `Shopping/Tobacco, Recreation/Tobacco`
- Dating: `Society/Relationships/Dating,Society/Relationships/Cyber_Relationships,Regional/Europe/UnitedKingdom/Society_and_Culture/Gay,_Lesbian,_and_Bisexual/Relationships`
- Social Networking: `Computers/Internet/On_the_Web/Online_Communities/Social_Networking,,Kids_and_Teens/People_and_Society/OnlineCommunities`
- Gambling: `Gambling`
- File-sharing: `Computers/Software/Internet/File Sharing`
- Violence: No category - assume that this is covered by `Adult`
- Gaming: `Games`
- Malware and Hacking: `Computers/Hacking`