RESEARCH

An Investigation of the Performance of UK ISP Web Filters

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

[*]Correspondence: richard@openrightsgroup.org
[2]Open Rights Group, Free Word Centre, 60 Farringdon Road, EC1R 3GA London, UK
Full list of author information is available at the end of the article

Abstract
To do

Keywords: sample; article; author

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 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**: How reliable are the filters at blocking content, in terms of both overblocking and underblocking across different categories of sites? And are there certain categories of web sites that are error prone?
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

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

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

[2] A non-profit UK-based organisation who campaign for and work to promote digital rights

[3] `https://www.blocked.org.uk/`

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

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 how and whether 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. filter request, modify response); 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: "*...this 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 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

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

(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]

The study of web filtering and its mechanisms transcends various layers include 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 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 that 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*'.

Despite the wide body of work detection filtering approaches and their usage across various countries, we could only find one piece of work that empirically characterised web filter accuracy. The work in question, by Stark [13], used a sample of URLs categorised as either adult or clean, and passed these URLs through various home PC web filters (e.g. McAfee, CyberPatrol) looking for: (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 ranged from minimum of 6.2% to 43.4% underblocking percentages for various filters, and 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 of UK ISP and MSP web filters, nor their uptake - aside from the Ofcom Internet Safety report in 2014.[6]

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

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

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

-Explain which internet service providers are included –Different blocking settings used and analysed (e.g. BT moderate, BT strict, etc.) -Explain what —SPs use which companies services

Analysed Internet Service Providers

Broadband Providers

-BT -Plusnet -Sky -TalkTalk -VirginMedia

Mobile Providers

-EE -O2 -T-Mobile -VirginMobile -Vodafone

Blocking Approaches

-BT: –provides a system known as 'BT-parent controls'[7] which uses DNS-based blocking of URLs –uses site categorisation information from Nominum –Provides three levels of filtering once controls are 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 setting's content and 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 –provides a system known as 'Plusnet Protect'

-Sky –provides a system known as 'Sky Broadband Shield' which also uses DNS-based blocking of URLs –uses site categorisation information provided Symantec and their Rulespace Web Content categorisation system[8] –Also offers three levels of categorisation: (i) 18 which 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 Homesafe –Uses DPI to examine URLs being visited by users. Sites blocked using DNS-spoofing –Includes setting of '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", "Weapons and Violence".

-VirginMedia: provides a system known as 'Web Safe'[9] which is a DNS-based system that matches requested URLs with known blocked URLs in a DNS-lookup table. –uses site categorisation information from Nominum –Data was not immediately available of what VirginMedia blocks, therefore used the OFCOM Internet Safety Measures report.

Good overview of which categories the filters cover is included in the OFCOM Internet Safety Measures report from 2014.[10]

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

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

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

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

Monitoring Web Filters

-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 -Show the distribution of requests per day per ISP filter

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

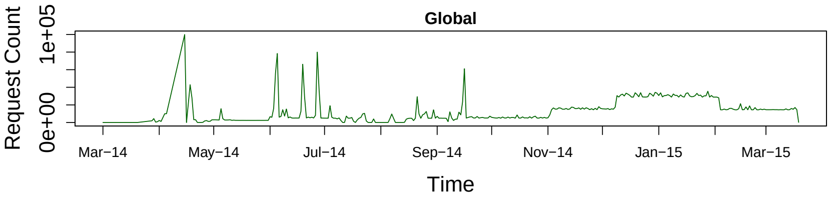


Figure 2 Number of URL requests made per broadband ISP filter

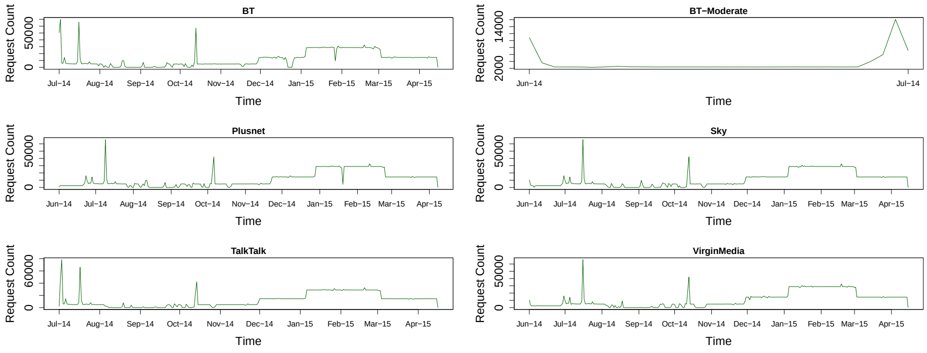
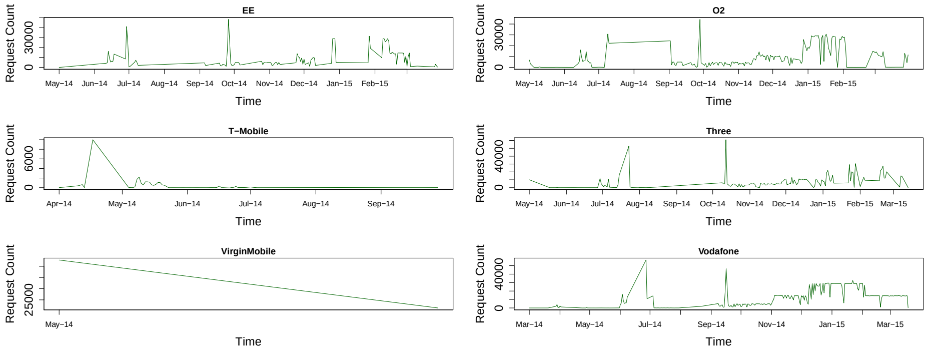


Figure 3 Number of URL requests made per mobile ISP filter



Mobile ISP findings: -Dropping VirginMobile from the analysis as there was a problem with the data -Can only analyse T-Mobile up to the end of September 2014

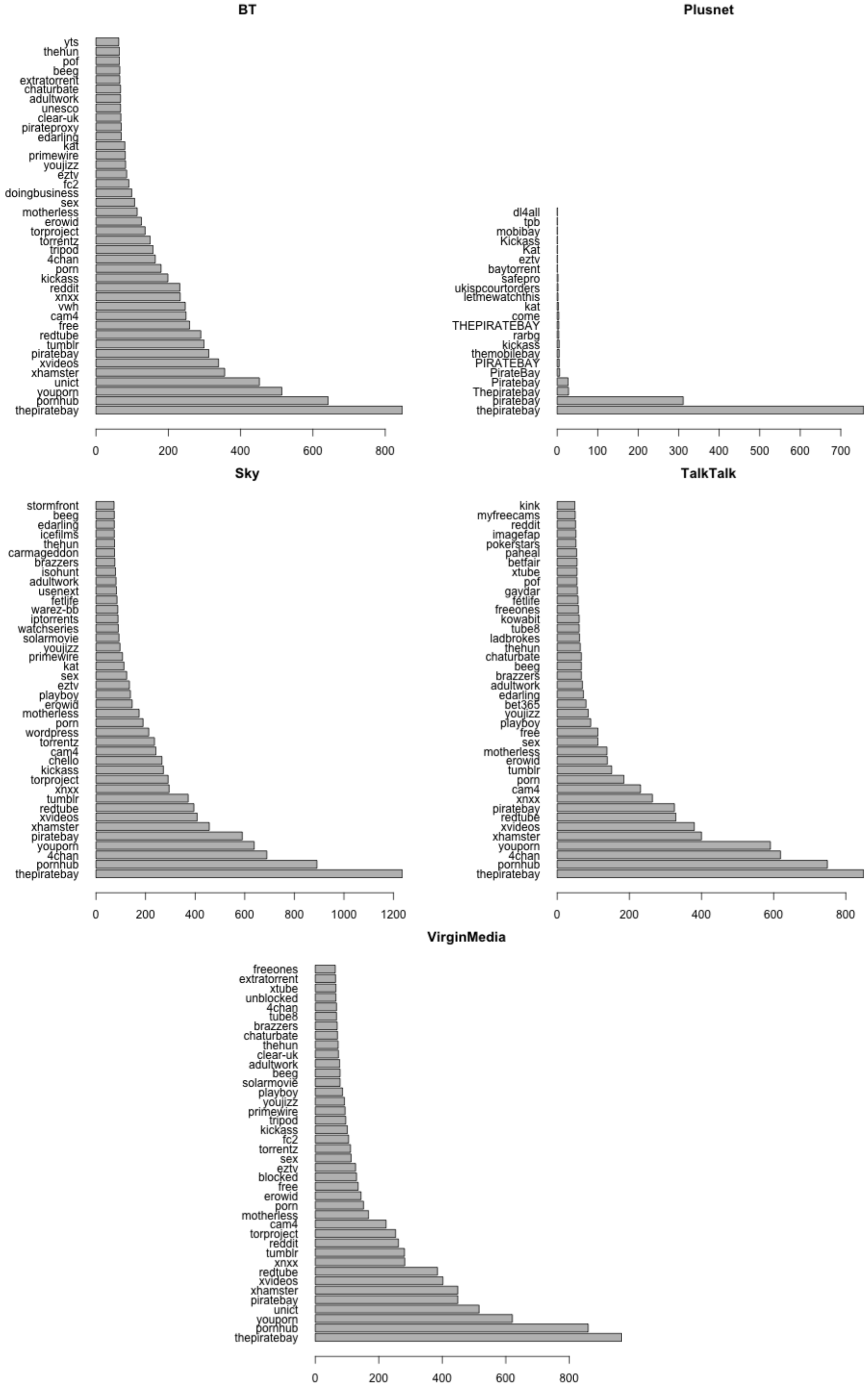
Dropped filters: -BT-Moderate -VirginMobile -T-Mobile

Blocked Content

Blocked Domains

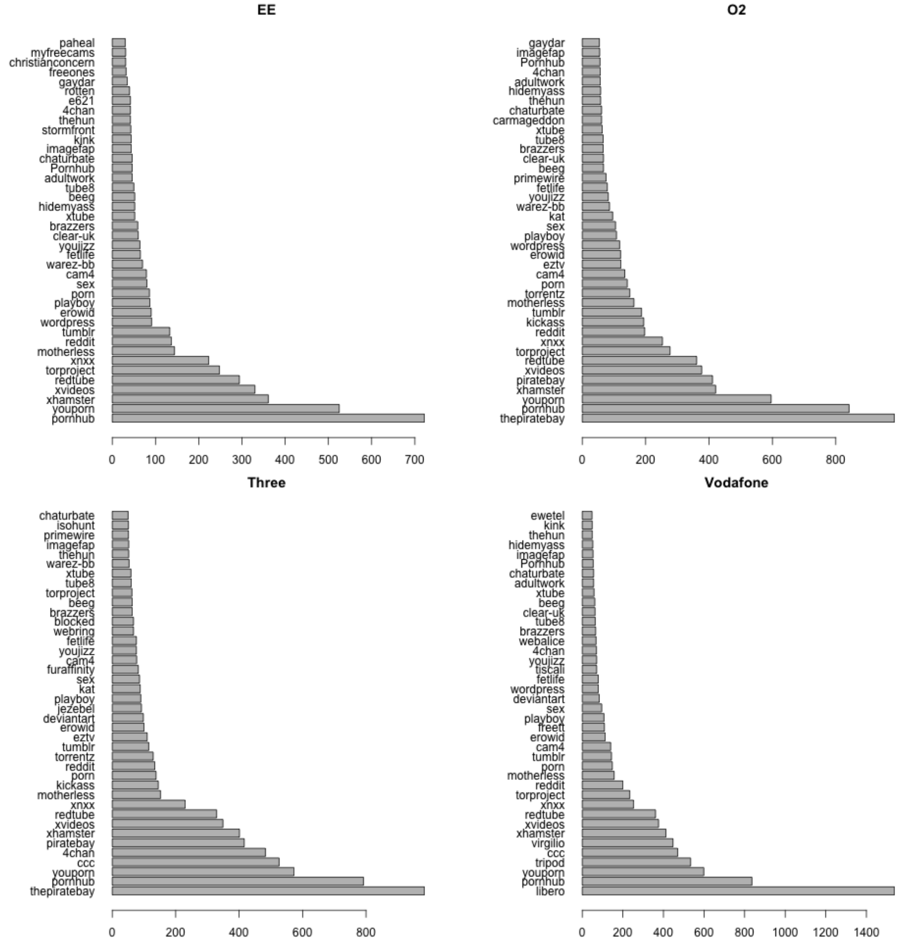
-Report on distribution of domains that are blocked so far to date (ignoring changes)

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



To do: -Add analysis of wordpress, tumblr, reddit, and livejournal sites

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



Wordpress Examples: -URL: http://beeractivist.wordpress.com — Submitted: 2014-11-30 21:13:51 — NetworkName: TalkTalk — Status: blocked -URL: http://tantracore.wordpress.com — Submitted: 2014-11-30 22:13:28 — NetworkName: Sky — Status: blocked -URL: http://garsai.wordpress.com — Submitted: 2014-11-30 23:02:46 — NetworkName: TalkTalk — Status: blocked -URL: http://toysoldier.wordpress.com — Submitted: 2014-07-03 15:03:15 — NetworkName: O2 — Status: blocked –This site is a support site for men who have been abused (also blocked by Sky) -URL: http://www.heyartist.wordpress.com — Submitted: 2014-11-30 20:56:34 — NetworkName: Sky — Status: blocked –Site promoting art as a support mechanism for enhancing wellbeing

Tumblr Examples: -URL: http://atlasofprejudice.tumblr.com — Submitted: 2014-05-14 01:12:57 — NetworkName: TalkTalk Strict — Status: blocked –Showing examples of maps that demonstrate the prejudices that countries have -URL: http://azurelunatic.tumblr.com/post/18654147576/ive-been-forced-to-explain-homosexuality-to-my — Submitted: 2014-05-27 21:27:00 — NetworkName: Vodafone — Status: blocked –Example of blocking a site as it contains an explanation of why some-one is gay (potential prejudice here). Also blocked by O2, TalkTalk, and BT -URL: http://thusly.tumblr.com — Submitted: 2014-07-02 13:08:44 — NetworkName: EE — Status: blocked –Also blocked by BT, Sky, O2, Vodafone -URL: http://tldrwikipedia.tumblr.com — Submitted: 2014-05-14 01:13:11 — NetworkName: TalkTalk — Status: blocked -URL: http://notalkingplz.tumblr.com — Submitted: 2014-07-02 14:22:20 — NetworkName: O2 — Status: blocked –Also blocked on: Sky, Vodafone, EE

Reddit Examples: -Largely blocking http://www.reddit.com/r/nsfw, http://www.reddit.com/r/nsfl, and http://reddit.com/r/porn, all of which are sub-reddits containing adult content -URL: http://www.reddit.cm — Submitted: 2014-07-02 10:40:30 — NetworkName: Sky — Status: blocked -URL: http://reddit.com/r/creepypms — Submitted: 2014-07-05 20:10:31 — NetworkName: EE — Status: blocked –Subreddit sharing creepy private messages that people have received. Not necessarily adult content, and definitely not pornography.

Livejournal: -All appear to the sites of Russian sites (e.g. http://limonov-eduard.livejournal.com/)

-URL: http://community.livejournal.com/asi/ — Submitted: 2014-11-30 21:11:19 — NetworkName: Sky — Status: blocked –Anorexia and self-harm support community site. Contains posts from people explaining their afflictions and getting support from other people.

-URL: http://beer-retard.livejournal.com — Submitted: 2014-11-30 21:13:51 — NetworkName: TalkTalk — Status: blocked –Blocked due to discussing/containing information about alcohol/beer?

-URL: http://urban-decay.livejournal.com — Submitted: 2014-11-30 21:14:59 — NetworkName: Sky — Status: blocked –Also blocked by Three and O2

-URL: http://ercasse-ainince.livejournal.com/30230.html — Submitted: 2014-11-30 20:51:44 — NetworkName: Sky — Status: blocked –Article about films that have been out for a long time

-Largely blocking pornography live journal pages too

Blocked Site Categories

-Report on the distribution of categories of sites that are blocked –Explain the categorisation system used, and the various levels of categories –What is the coverage of URLs in the DMOZ categorisation system? Report on the % covered in the system

Given the DMOZ categorisation system and the use of ODP categories, we can use the hierarchy of the category taxonomy to only focus up to a specific depth; thereby restricting the categories to only depth of d.

-Not showing Plusnet as it blocks file-sharing category sites

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

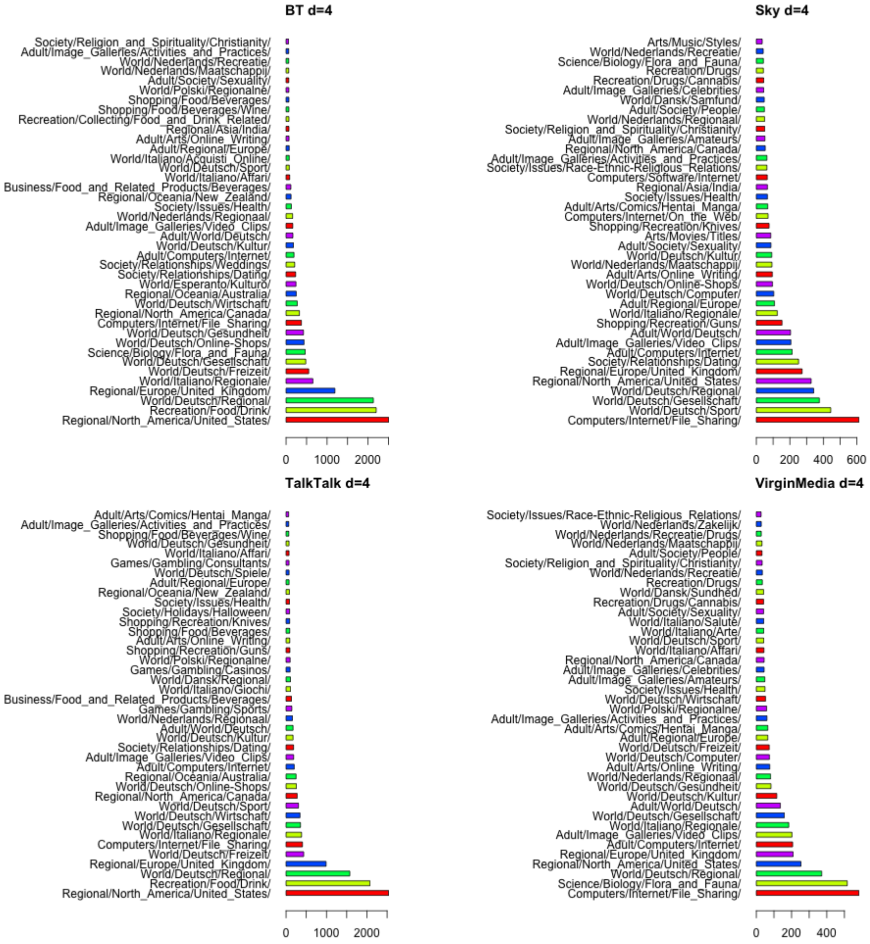
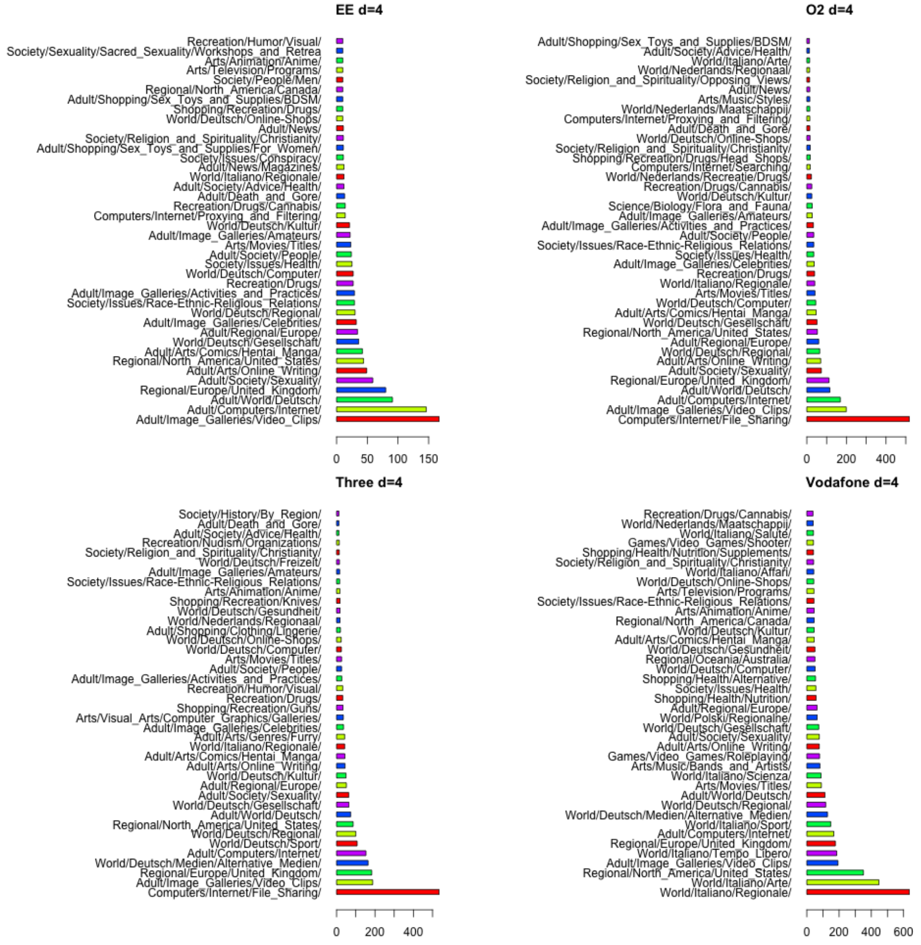


Figure 7 Distributions of blocked level-4 categories for mobile ISPs.



Gauging Filter Accuracy

-Report on reverse engineering the accuracy of the filters –Need to consider how to collect certain porrnographic examples here and push them to the blocked.org.uk queue –Report on the class balance between blocked vs not-blocked content –Can only consider URLs within the system? –Do the findings suggest a certain problem with the mechanisms used to classify urls?

Pseudo-Classifiers for ISP Filters

-Induce a pseudo-classifier for each ISP filter setting: –Specify which categories of content should be blocked by certain filter –Add table to explain this –Coded blocked topics into DMOZ categories to identify what the pseudo classifier should block. –Our approach is to always go as conservative as possible: i.e. if unsure about whether a category should be blocked, then block it. –Remove file-sharing for now, as this is assessed on a case-by-case basis and with court orders —I.e. blocking all of Computers: Software: Internet: File Sharing would lead to studies on file-sharing being blocked –Hence: block everything under Adult –Also block everything under Computers/Hacking

-BT: should block... –pornography (Adult/Image Galleries + Adult/Video Clips) –obscene and tasteless (Adult/Death and Gore) –hate and self-harm (no DMOZ category) –drugs (Recreation/Drugs) –alcohol (Recreation/Food/Drink/Drinking + Health/Specific Substances/Alcoholic Beverages) –tobacco (Shopping/Tobacco + Recreation/Tobacco) –dating (Society/Relationships/Dating, Society/Relationships/Cyber_Relationships)

-Sky: should block... –malware sites (Computers/Security/Malicious_Software/Spyware_and_Adware) –cyber-bullying (no cat on this, generally there are advice pages though) –pornography (Adult) –suicide and self-harm (no cat) –drugs (Recreation/Drugs) –dating (Society/Relationships/Dating) –social networking (Computers/Internet/On_the_Web/Online_Communities Kids_and_Teens/People_and_Society/Online Communities) –online gaming (Games/Online)

-TalkTalk: should block... –dating (Society/Relationships/Dating) –drugs (Recreation/Drugs) –alcohol (Recreation/Food/Drink/Drinking + Health/Specific Substances/Alcoholic Beverages) –tobacco (Shopping/Tobacco + Recreation/Tobacco) –File Sharing Sites (Computers/Software/Internet/Clients/File Sharing) –Gambling (Gamling) –online gaming (Games) –Pornography (Adult) –social networking (Computers/Internet/On_the_Web/Online_Communities/Social_Networking, Kids_and_Teens/People_and_Society/ Communities) –Suicide and Self-Harm (no cat) –Weapons and Violence (Adult)

VirginMedia: should block... –Crime, Violence, and Hate: (Adult) –Drugs (Recreation/Drugs/Cannabis + Recreation/Drugs/Psychedelics) –File Sharing Sites (Computers/Software/Internet/Clients/File Sharing) –Pornography (Adult) –Suicide and Self-harm (no category for this in DMOZ)

-EE, O2, and Three: should block...

–18 works are for adults and can contain strong issues such as: very strong violence, frequent strong language (e.g. 'f***') and / or very strong language (e.g. 'c***'), strong portrayals of sexual activity, scenes of sexual violence, strong horror, strong blood and gore, real sex (in some circumstances), discriminatory language and behaviour

-Vodafone: should block... –Our content control prevents access to 18-rated content on Vodafone live! (mobile internet) and blocks access to 18-rated websites, un-moderated chat rooms and listed child abuse sites.

From the DMOZ web site: Generally the Adult category includes sites whose dominant theme is either: -To appeal to the prurient interest in sex without any serious literary, artistic, political, or scientific value -The depiction or description of nudity, including sexual or excretory activities or organs in a lascivious way -The depiction or description of sexually explicit conduct in a lascivious way (e.g. for entertainment purposes)

Judging Filter Accuracy

Explain gauging filter accuracy using existing measures from classification literature. -Parallelisation of the accuracy measurement -Explain the role of Spark in this to distribute the work load -Explain the location of the code and how to run this -Explain how we calculate accuracy based on which requests were blocked and which were not blocked - we have to do this as a URL can go from blocked to unblocked and vice-versa –Question: how often is a given URL periodically tested for a block?

Compare results to that of [13]

General Accuracy

Table 1 Accuracy levels of ISP and Mobile Providers' Web Filters derived using the DMOZ categories that should have been blocked by each filter and the categories of URLs that were actually blocked.

	Precision	Recall	FPR	MCC	F1
BT	0.032	0.613	0.012	0.138	0.061
Sky	0.088	0.370	0.003	0.179	0.142
TalkTalk	0.078	0.073	0.009	0.066	0.075
VirginMedia	0.050	0.508	0.003	0.159	0.091
EE	0.189	0.635	0.002	0.346	0.291
O2	0.136	0.697	0.002	0.307	0.227
Three	0.108	0.631	0.004	0.260	0.185
Vodafone	0.044	0.564	0.004	0.156	0.081

Table 2 Accuracy levels after filtering out sites from the World category subtree.

	Precision	Recall	FPR	MCC	F1
BT	0.066	0.612	0.010	0.198	0.119
Sky	0.163	0.372	0.003	0.245	0.227
TalkTalk	0.145	0.072	0.008	0.091	0.097
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

Table 3 Accuracy levels after filtering out sites from the World category subtree and controlling for breweries and other alcohol related sites.

	Precision	Recall	FPR	MCC	F1
BT	0.335	0.726	0.007	0.490	0.459
Sky	0.163	0.372	0.003	0.245	0.227
TalkTalk	0.422	0.176	0.006	0.262	0.248
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

Qualitative Examples of Blocks: -BT Block: http://www.lgbtquitsmoking.com/ (site to help people stop smoking).

Table 4 Accuracy levels after applying keyword filtering for additional alcohol-related categories (e.g. brewery, wineries).

	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

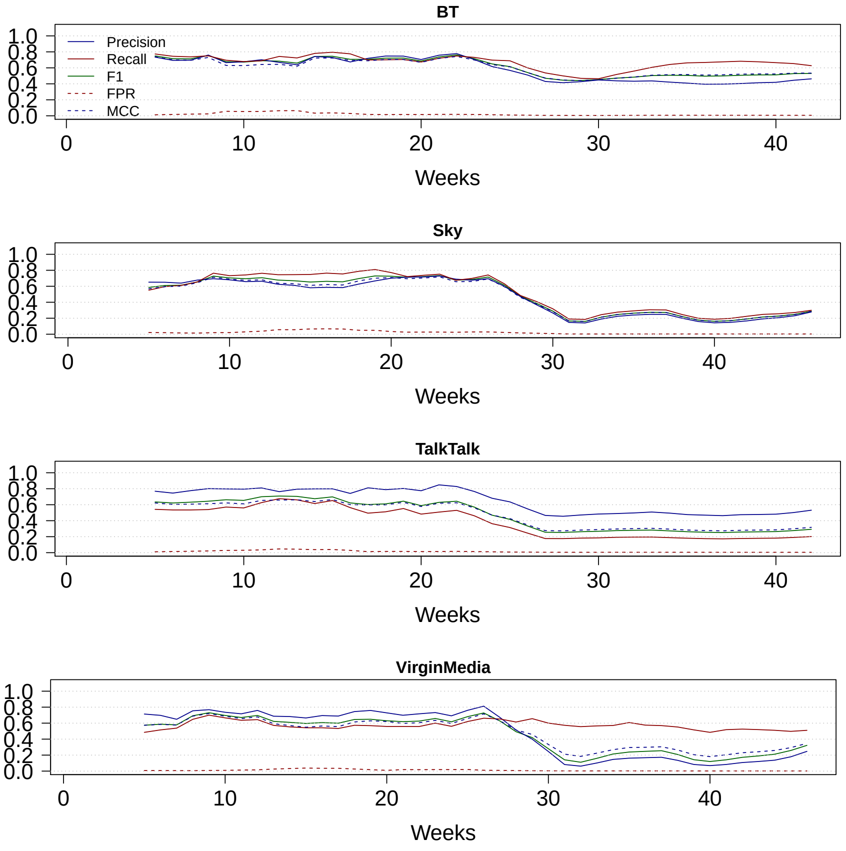
BT seem to be blocking tattoo web sites too:

We have uploaded the collection of sites which are false positives and false negatives to the github repo.[11]

Accuracy over Time

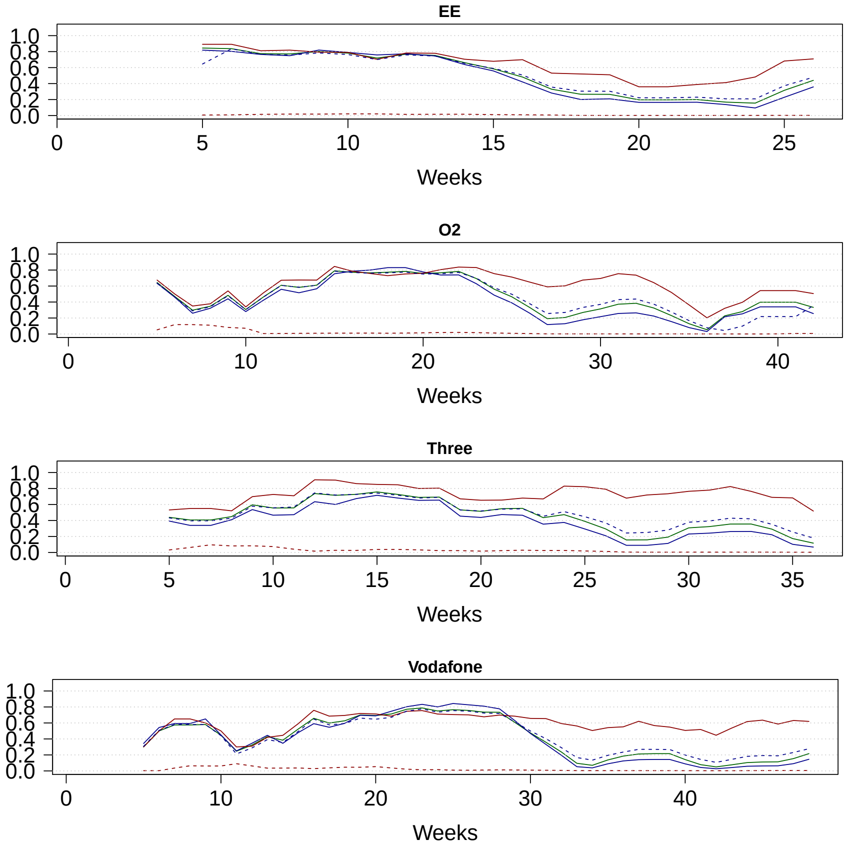
How has accuracy evolved over time? –Discrete time analysis of the accuracy levels - use weekly bins (Bayesian model?) Can we forecast accuracy?

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.



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

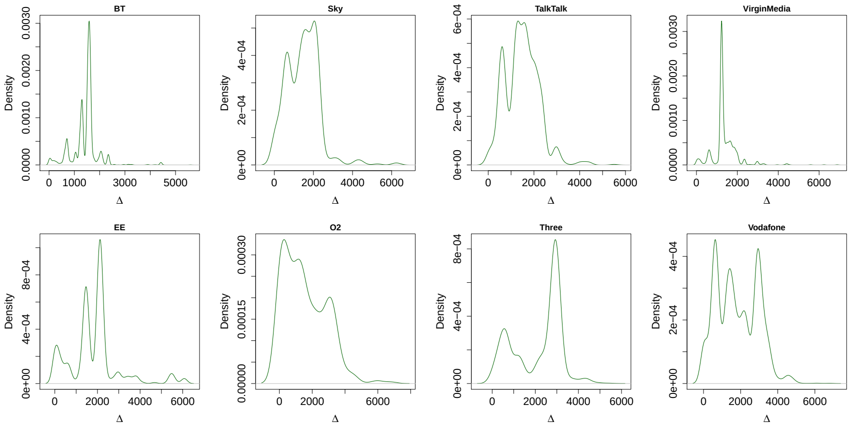
Figure 9 Accuracy of Mobile ISPs' filters over time. ARIMA(0,0,5) plot of the accuracy measures: precision, recall, f-measure (F1), false positive rate (FPR), and the Matthews' Correlation Coefficient. Weeks are from the start of the ISP-specific filter logging period.



Time to Correction

-Report on how long it takes each ISP to fix their blocked content -Show the delta distribution of each provider –Fit a distribution to the delta-function: Poisson?

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



Study Limitations

Measurement of Blocks

-Defend the approach of analysing which requests were fulfilled - this can contain duplicate URLs (some of which were blocked, and some of which were not) –This can lead to repeated URLs in the lists of false/true positives and negatives –We counteract this by using sets to restrict each URL to one occurrence per set

-Explain possible limitations with the actual probe system itself.

Limitations of Pseudo-Classifiers

Limitations of this approach: -Relies on the classification of sites within DMOZ as being correct -Coverage of the DMOZ categories - as this is manually curated we only cover a % of the URLs in total that have been aligned with categories

-Use of DMOZ categories is not without errors: –E.g. the URL http://www.vin-gastronomie.com/ is not classed as should be blocked in the gold standard as its category is "World: Francais: Regional: Europe: France: Regions: Haute-Normandie: Eure: Commerce et economie: Gastronomie et alimentation', however the page describes wine brands

-Potential improvements: –Classifying content of the page to mine topics discussed therein - i.e. basic semantic analysis of the content –Filtering out categories of sites which may introduce noise into the results, and not counting them at all (e.g. those related to gastronomy).

Findings and Implications

Conclusions and Future Work

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)