

RESEARCH

# An Investigation of the Performance of UK ISP Web Filters

Matthew Rowe<sup>1,2,3,4</sup> and Richard King<sup>1,2,\*</sup>

\*Correspondence:

richard@openrightsgroup.org

<sup>1,2</sup> 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

Context/Problem -Introduction of UK ISP Web Filters -Questions over: –the need for this (cite ofcom survey states) –their accuracy, given that the ISPs use third-party systems or categorisations –how they can be held to account –who is responsible for their oversight (this is left vague)

Motivation: -If we can study and therefore understand how web filters function, then we can also gauge their accuracy and in turn perceive how effective they are. We can also understand what the by-products are of this censorship

Research Questions: -Can we understand how UK ISP Web Filters function, and how accurate they are? -Are they reliable at blocking content, in terms of both overblocking and underblocking? And are there certain categories of web sites that are error prone? -How long does it take an ISP to fix an error?

This paper presents the first systematic study of UK ISPs' web filters, how they function, their accuracy, and their failings. We take a data science perspective and use data extracted from ISP filter probes

We adopt a macro-to-macro analysis of the data to first examine

## Related Work

-Explain existing studies of web censorship -Explain how the UK approach is different as it is via UK ISPs and therefore pushed out to the private sector

## Studied Internet 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’<sup>[1]</sup> 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<sup>[2]</sup> –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’<sup>[3]</sup> 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.<sup>[4]</sup>

---

<sup>[1]</sup><http://www.productsandservices.bt.com/products/manage-broadband-extras/>

<sup>[2]</sup><http://www.symantec.com/page.jsp?id=rulespace>

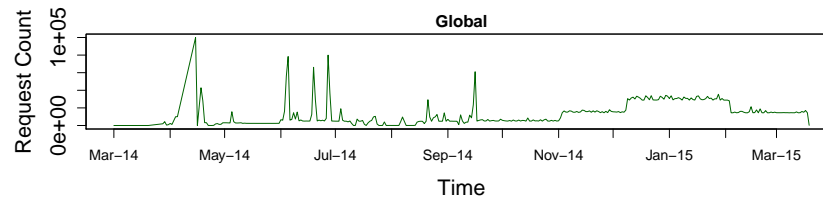
<sup>[3]</sup><http://my.virginmedia.com/my-apps/websafe.html>

<sup>[4]</sup>[http://stakeholders.ofcom.org.uk/binaries/internet/internet\\_safety\\_measures\\_2.pdf](http://stakeholders.ofcom.org.uk/binaries/internet/internet_safety_measures_2.pdf)

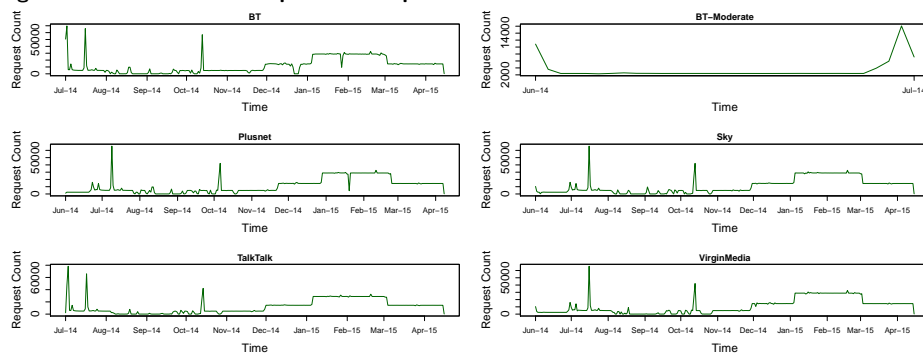
## Monitoring Web Filters

-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

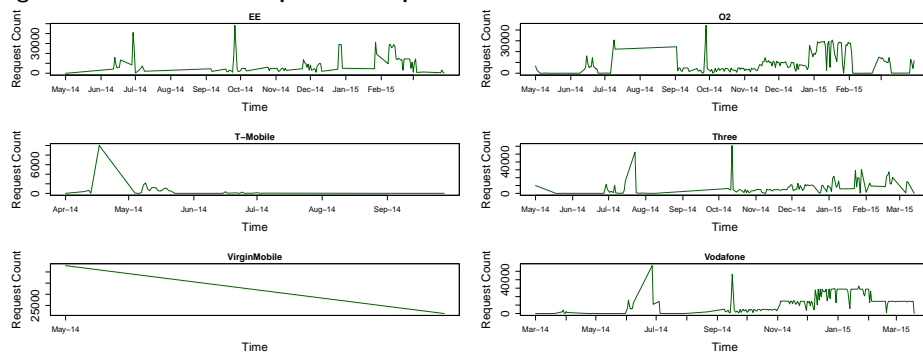
**Figure 1** Number of URL requests made over time since the beginning of the project.



**Figure 2** Number of URL requests made per broadband ISP filter



**Figure 3** Number of URL requests made per mobile ISP filter



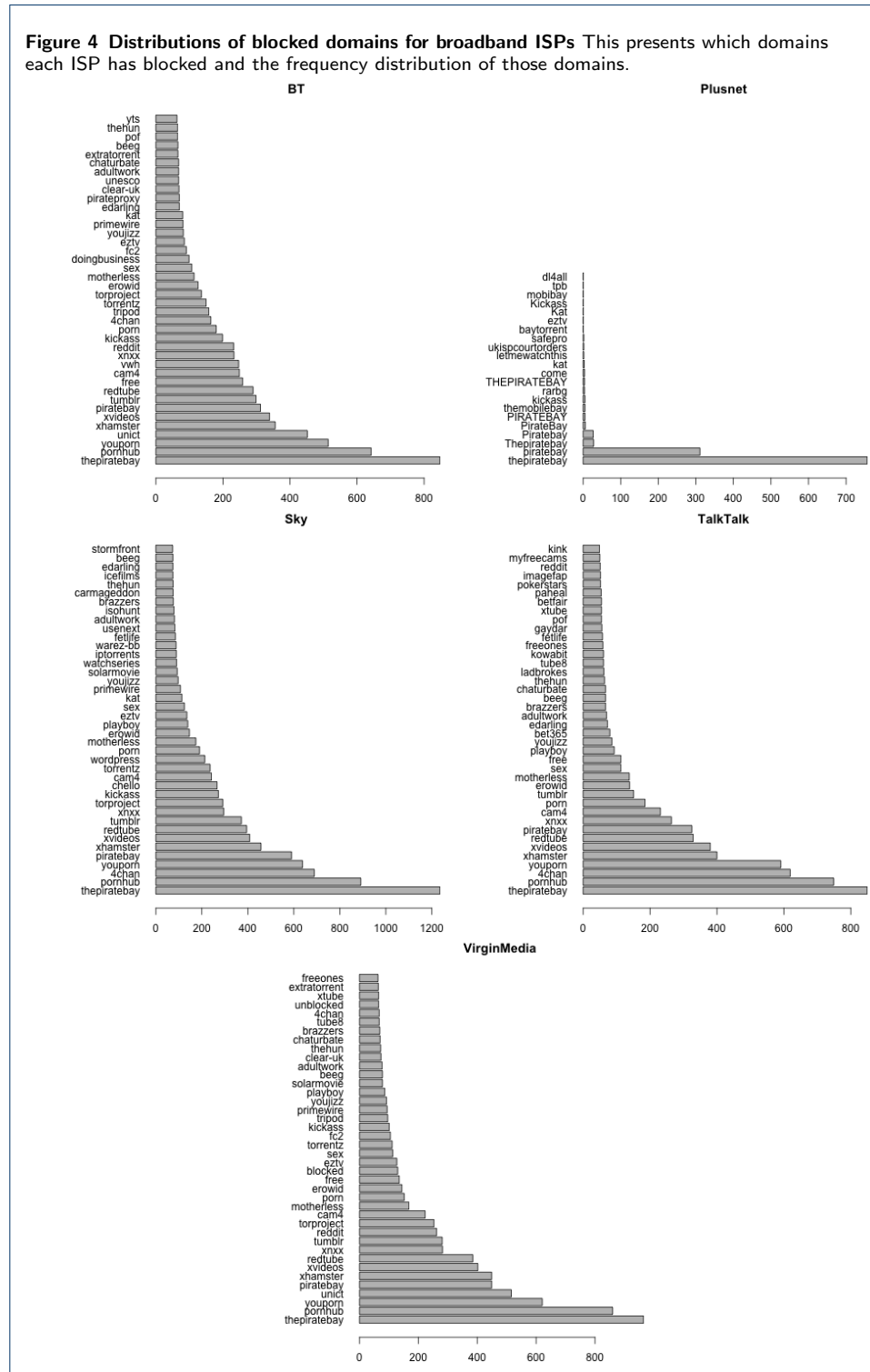
Mobile ISP findings: -Dropping VirginMobile from the analysis as there was a problem with the data -Can only analyse T-Mobile up to the end of September 2014

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

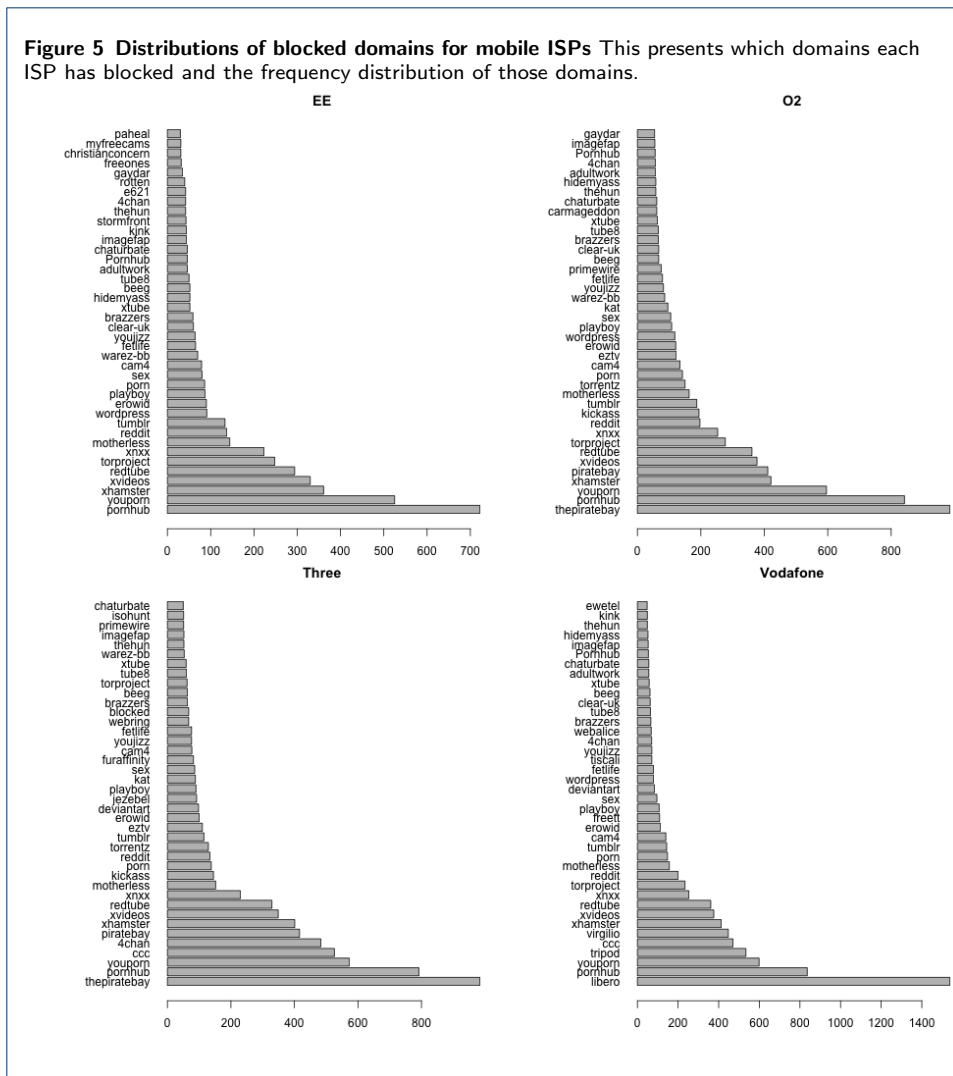
## Blocked Content

### Blocked Domains

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



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



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 — Network-  
 Name: EE — Status: blocked —Also blocked by BT, Sky, O2, Vodafone -URL:  
<http://tldrwikipedia.tumblr.com> — Submitted: 2014-05-14 01:13:11 — Network-  
 Name: TalkTalk — Status: blocked -URL: <http://notalkingplz.tumblr.com> — Sub-  
 mitted: 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/](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 commu-  
 nity 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 cate-  
 gorisation 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.

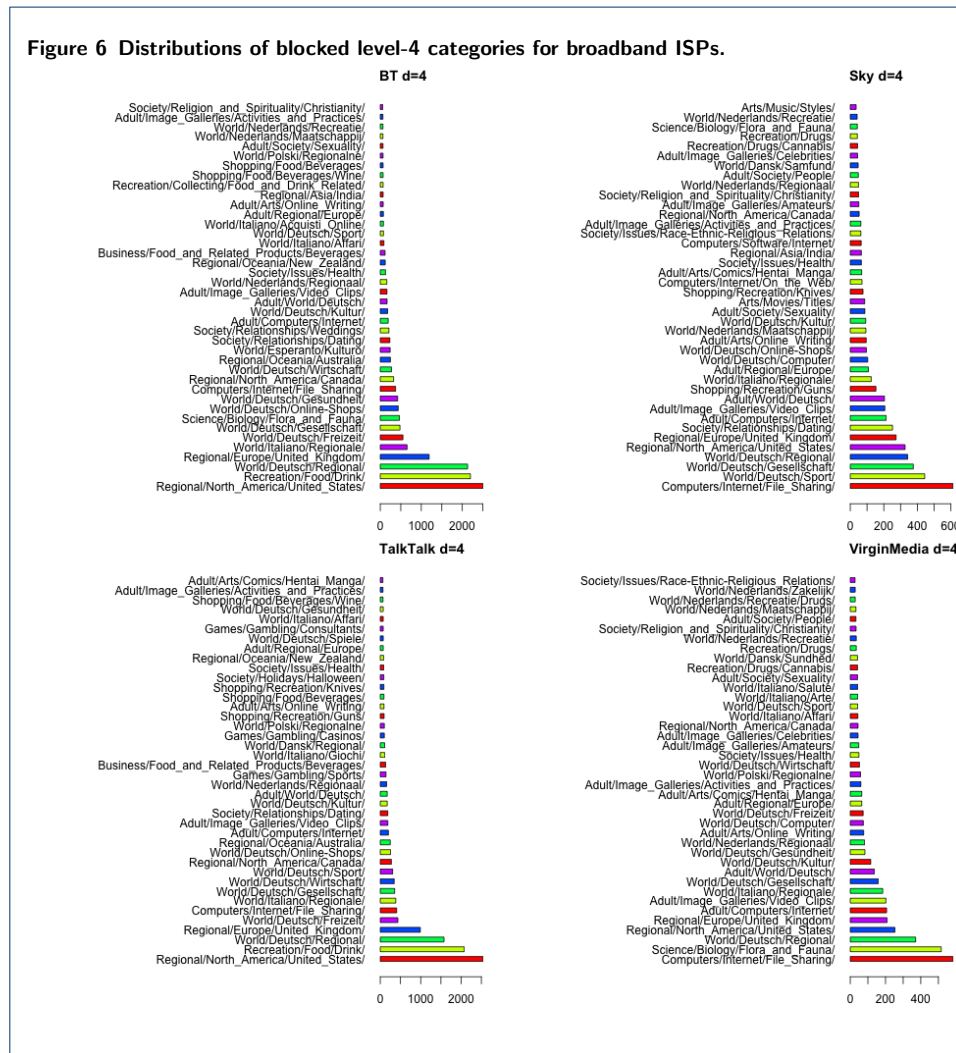
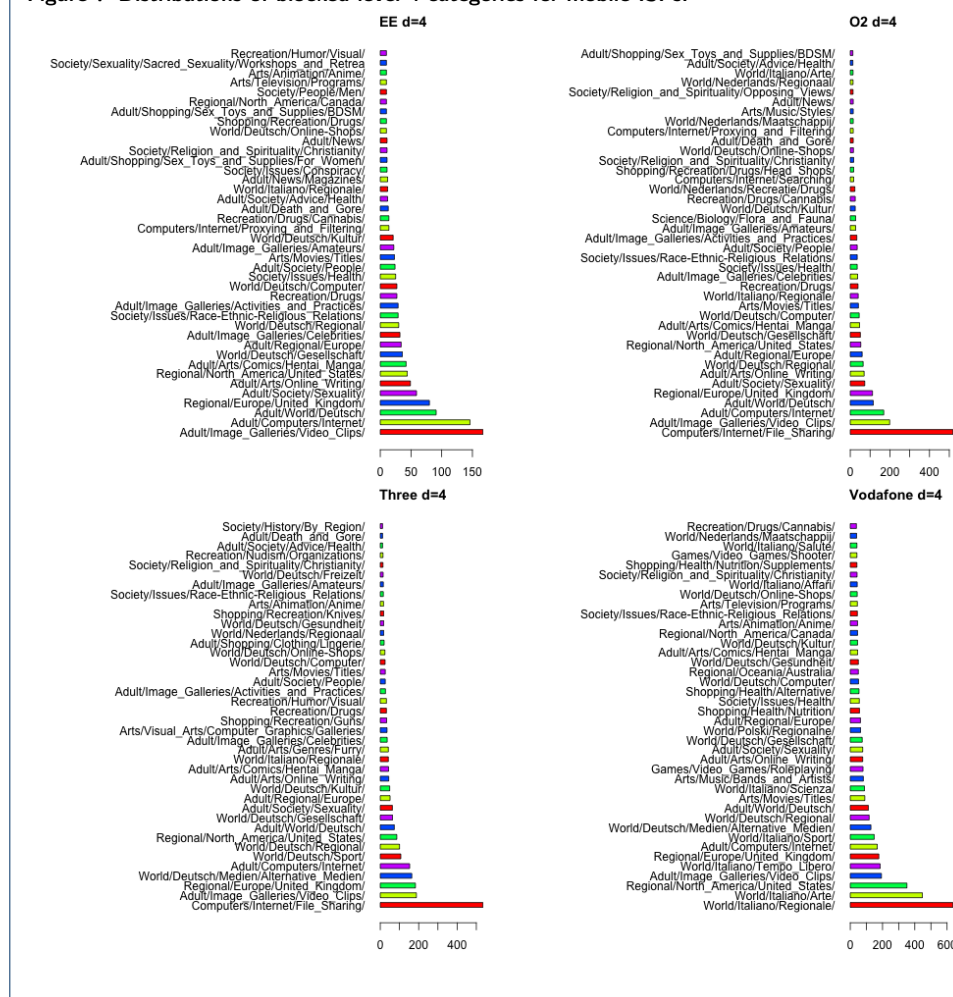


Figure 7 Distributions of blocked level-4 categories for mobile ISPs.





## Gauging Filter Accuracy

–Report on reverse engineering the accuracy of the filters –Need to consider how to collect certain pornographic 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?

### General Accuracy

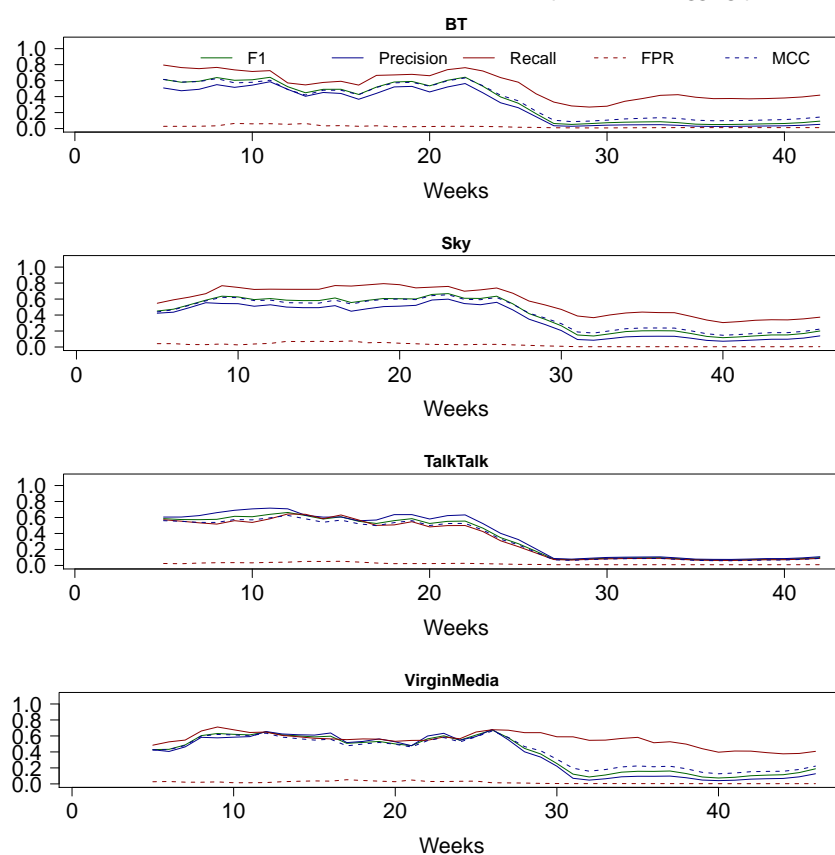
**Table 1** Accuracy levels of ISP and Mobile Providers' Web Filters derived using the DMOZ categories that should have been blocked by each filter and the categories of URLs that were actually blocked.

	Precision	Recall	FPR	MCC	F1
BT	0.083	0.550	0.012	0.210	0.144
Sky	0.217	0.556	0.004	0.345	0.312
TalkTalk	0.144	0.121	0.009	0.121	0.131
VirginMedia	0.185	0.611	0.003	0.335	0.284
EE	0.445	0.750	0.002	0.576	0.559
O2	0.324	0.794	0.002	0.507	0.461
Three	0.256	0.820	0.005	0.456	0.390
Vodafone	0.142	0.684	0.004	0.311	0.236

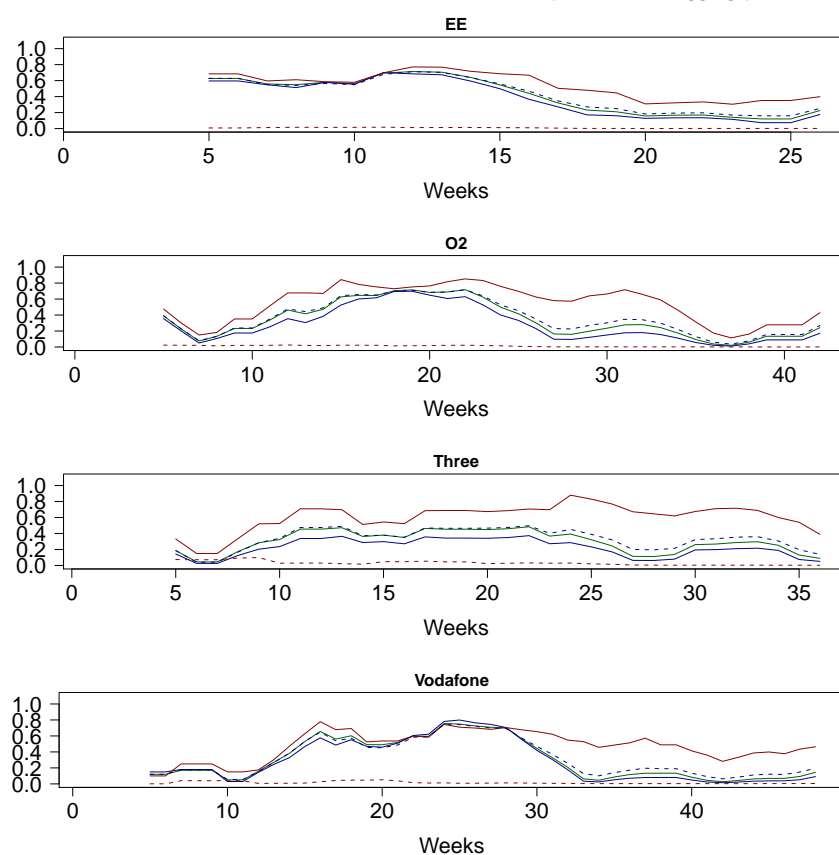
### Accuracy over Time

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.

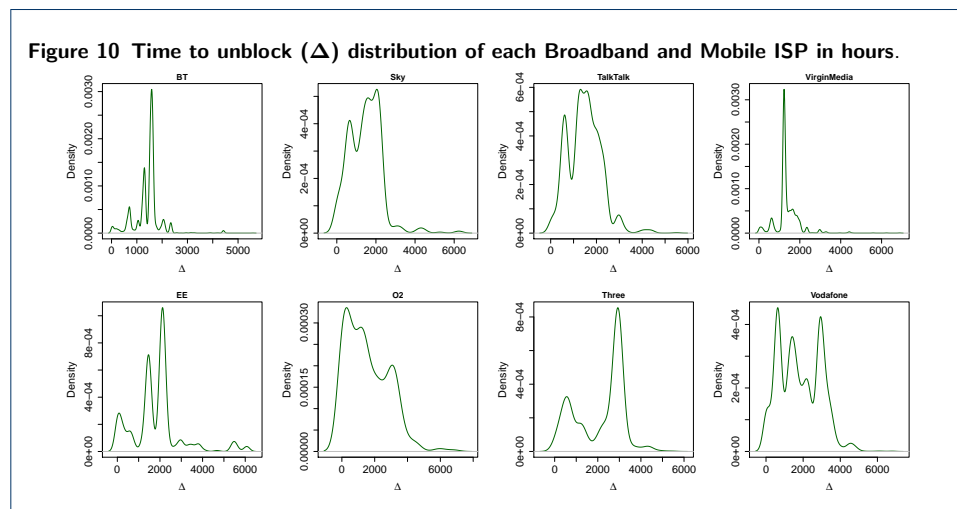


**Figure 9 Accuracy of Mobile ISPs' filters over time.** ARIMA(0,0,5) plot of the accuracy measures: precision, recall, f-measure (F1), false positive rate (FPR), and the Matthews' Correlation Coefficient. Weeks are from the start of the ISP-specific filter logging period.



## 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?



## 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

## 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