Internet Censorship - The State of Censoring in the United Kingdom

1st Anna Franziska Bothe
Information Science Faculty
Humboldt Universität zu Berlin
Berlin, Germany
botheann@hu-berlin.de

2nd Jan Krol
Information Science Faculty
Humboldt Universität zu Berlin
Berlin, Germany
kroljan@hu-berlin.de

3rd Louis Kiesewetter
Information Science Faculty
Humboldt Universität zu Berlin
Berlin, Germany
kiesewel@hu-berlin.de

***Abstract*—Internet censorship is not a topic exclusive to countries with oppressive regimes. Over the last two decades, the government of the United Kingdom (UK) has introduced a number of policies to restrict access to certain types of content on the internet. In this work the goals and legal foundations for these policies are described. The extent of censorship in the UK is illustrated by analysing data from internet censorship measurements, provided by OONI and the Blocked project. Furthermore, we outline the problem of legitimate websites being falsely blocked as a result of the enforced policies. In spite of the amount of applied internet censoring rules in the UK, the applied blocking mechanisms can still be relatively easy circumvented. This raises the question whether the social costs of a restriction in personal freedom for all citizens can be justified by a hampered access to harmful and indecent content on the internet.**

***Index Terms*—internet censorship, united kingdom, privacy, OONI, blocked, filtering, clustering**

I. Introduction

Globally, the use of the internet is steadily rising. With less than 16% of people worldwide being regular users just 15 years ago, the number of people connected to the web is nowadays estimated to be over 50% and steadily rising still [1]. This continuing trend can be attributed to improving living standards and connectivity in developing countries, but is nonetheless also observable in developed nations, where close to 87 % of the population nowadays regularly makes use of the internet [1]. Adding to the importance of the internet for a nation's communication is the growing impact on the GDP the internet has [2]. This in turn makes the internet a tool of necessity for many businesses and ever increasing connectivity an inevitability. However, with the internet offering a largely unmonitored exchange of information, governments are growing ever more intent on restricting and monitoring internet access. One of the most widely known governments to do so is the Chinese one; however, internet censorship is also present in democratic first-world countries [3]. The United Kingdom is currently filtering web traffic for a majority of its citizens, with relatively few attention being given to this form of censorship in recent research.

This paper looks at the current websphere in the United Kingdom and, through data collection and analysis, highlights the extent to which the British internet is being censored. It gives an overview of the filtering and monitoring present in the UK internet as well as its legal basis and technical implementations.

II. Overview of Censorship

As an introduction to censorship an overview about censoring in general is given. Hence, a definition of the term censorship is given, the legal foundation of the UK as well as comparable countries is examined. Afterwards, the historic background and development are described and characteristics of censorship as well as the drawbacks and benefits of censoring are discussed.

A. Definition of Censorship & its Actors

In general, the term censorship can be defined as "the suppression or alteration of speech or writing prior to publication in the interests of an alleged higher social good." [4].

Censorship cannot only be enacted by a government but also by public and private individuals, such as film studios for instance [5]. Private groups, for instance, can censor a magazine by blocking the entrance of the store which sells it. For the most part, those groups are harmless and act on a legal basis, but might become dangerous in extremes.

B. Legal foundation

The legal ground for censoring actions of private groups as well as the protection of governmental restrictions is provided by the freedom of speech – the United States for example have the first Amendment, while Germany has its Article 5 of the constitution (Art. 5 I, III GG) that guarantees the freedom of speech, expression and opinion [6] [7]. In contrast, free speech in the United Kingdom is only protected by the Human Rights Act from 1998, which is an attempt of incorporating the European Convention into the domestic law. It, however, includes a lot of exceptions [8]. Thus, freedom of speech in the UK is subject to limitations and censoring is much easier to legally apply.

Not all forms of government censorship are unconstitutional, e.g. in Germany, where some limitations are also given by the German law. There, the government is allowed to interfere and limit the freedom of speech via censorship as soon as the youth or the personal honor is endangered, cp. Art. 5 II GG. In the United Kingdom, the limitations on censorship

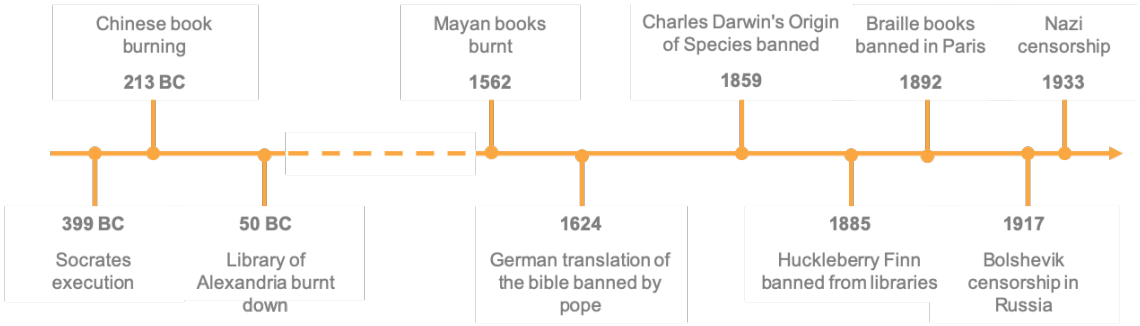

Figure 1. Overview over exemplary historic censorship actions.

are broader. As soon as it is "necessary in a democratic society, in the interests of national security, territorial disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary", censorship can legally be applied, Article 10 of the Human Rights Act.

C. History

Nowadays, all kinds of areas and media can be affected by censoring – arts, entertainment, literature, social media, speeches, music, pictures and every other forms of creative expression [9]. However, censoring content is not an invention of the 20th or 21st century.

Censoring has its origin in ancient societies. Rome as well as ancient Greece used censorship as a tool for maintaining political or moral values in society. In 399 BC, the first famous case of conviction due to not being in line with the censor establishment was the execution of Socrates. He was accused of having a bad influence on the youth and of being impious.

One hundred years later, China introduced their first censorship law. [10]

Besides introducing censorship laws, there are many other possibilities of censoring, such as the destruction of unwanted content. Book burnings also play a big role in the history of censorship. The first mass booking burning took place in China in 213 BC, followed by the fires in the library of Alexandria in 50 BC and 700 AD[1] (the cause is still unsolved), the Royal Library of the Samanid Dynasty in Turkey (1000) and some other book burnings e.g. in Mexico (1562) and the very well-known book burning of the Nazis in 1933.

A milder method of censoring written content was the proclamation of bans. Famous example of the book banning was the translation of the German bible in 1624, banning the origin of the species by Charles Darwin in 1859 or Huckleberry Finn by Mark Twain in 1885.

All described historic censoring activities were conducted by the government which is still the main actor of censoring any kind of content. As underlined by the examples, censoring is not a regional or culturally specific instrument. It happened and happens all around the world. The main reasons for censoring are conflicts to ideologies, beliefs, morals or simply Nazism. For instance, the book burning in China in 213 BC took place because the Chinese emperor Qin Shi Huang wanted the people to believe that the world started with his reign. [11]

All these reasons seem to be embedded in one main emotion: fear. It might be the fear of the unknown, fear of losing power and fear of losing control.

D. Characteristics

As outlined above, censorship can take on different shapes such as lawful bans of unwanted content or elimination of unwanted content. But any kind of censorship needs to carry certain characteristics.

It always censors at least one "protection area". Figure 2 shows that the protection area can be political, ethical, social, religious and/or military. It is called protection area because the government justifies the censorship through the protection of the people by censoring this area. For instance, military information is often censored to prevent opposing entities from gaining relevant security information. [12]

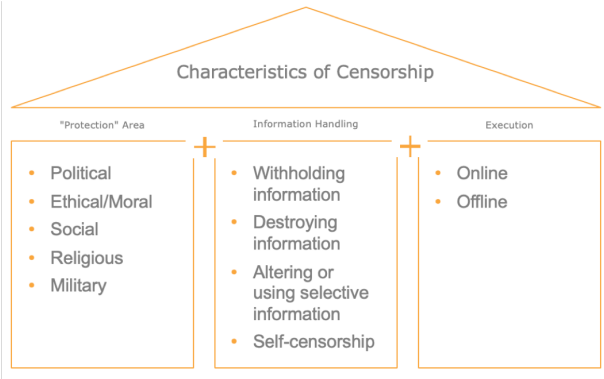

Figure 2. The three pillars of censorship.

Besides, censorship needs to choose an information "management" or, rather, handling. Information can either be withheld (as in the given military example), destroyed (e.g. book burning), altered or selected (e.g. announcing Corona has infected 100.000 people even though it has infected 500.000

[1]the following dates will all be AD.

people) or can also be self-censored (e.g. if a person is afraid of hurting someone or if the person is too afraid of the consequences that might arrive by publishing or sharing information with others) [13] [14].

Last but not least, without execution there is no censorship. The execution needs to be either online or offline but can also take action in both areas. A bullet point from every one of the pillars in Figure 2 has to be fulfilled in order to be characterized as censorship – several occurrences are also fairly possible.

E. Benefits vs Drawbacks

The word censorship is negatively connoted. However, the main goal of censorship is protection and security. Especially the youth needs to be protected from malicious content such as an execution of a human or disturbing pornography. [15] Censorship can also prevent the spread of (terrorist) propaganda and content that might recruit new followers. Furthermore, the access to illegal products such as weapons and drugs is more difficult. The introduction of laws against hacking strengthens the national security. Additionally, identity theft got harder by limiting the access of certain information. Hence, censorship can achieve a reduction of criminal activities. [16] But the question is: to what extent? And to what extent are the people willing to give up their right of speech and their liberty.

The reasonable extent of censorship is a trade-off between security and freedom. Increasing security by censoring content is following a thin line of serious interruptions and limitation of the freedom of speech/expression. Differentiating between hate speech and a negative expression or critics is often very difficult. [17] Moreover, the government decides which information is accessible which might lead to a biased flow of information and thus, the people are not able to get to the bottom of the truth [16]. Conclusively, censorship can hinder the citizen from forming an opinion based on the full information content that is available since they only have access to a part that the government perceives as appropriate information.

Striking about the enumerated advantages and disadvantages is that most of them gain importance in the internet. For example, identify theft as well as the distribution of disturbing content is easier, much faster and harder to prosecute online. As stated in the introduction of this paper, the focus will be on internet censorship with the main emphasis on the United Kingdom. Hence, the next chapter gives an overview of internet censorship in the UK.

III. Internet Censorship in the UK

Over a time period of more than two decades, the United Kingdom has implemented a plethora of internet censorship procedures across all networks and web communication devices [18]. In order to do so, the UK has made use of content filters and blocking mechanisms, with an ever expanding library of content being filtered. Facilitating further discussion of internet censorship in our context is a short definition of internet filtering and content blocking.

A. Definition of Internet Access

Internet access, as defined by the Cambridge Dictionary is the ability to connect to the internet [19]. However, connecting to the internet is not equivalent to interacting with services offered on the web or other means of internet-services. Only by accessing this web-content, censorship can actually be implemented and measures in network filtering can be implemented. As Techopedia specifies, a device is necessary in order to connect to the world wide web, this can be a personal computer (PC), an internet-capable smartphone or any other device with internet-capability [20]. No matter what device though, the United Kingdom's filtering measures are present across all end devices, regardless if they are private or public, such as library, university or school PCs.

There are many ways to obtain internet service, with broadband internet connections to the home and mobile networks being the most common in the UK [21]. Almost all methods of obtaining internet access as a private citizen are subject to censorship, with filtering methods being implemented across mobile networks, private internet connections as well as publicly available internet connections [22].

B. Internet Filtering vs Blocking

According to the National Coalition against censorship, internet filters are "software that prevents users of a computer from accessing certain websites" [23]. This definition is correct in that it captures one way in which access to the internet can be filtered. But filtering can happen in many more ways than just installed software preventing users of a predefined computer from their wanted access. An internet filter can also be an ad-blocking program installed by the webuser themself in order to prevent ads from being shown on their screen while accessing the content they want to see.

This voluntary and explicitly requested filtering is different from the filtering this paper looks at, which is in most cases involuntary and implemented at a point in the United Kingdom's central web infrastructure which cannot be interacted with or controlled by the user. Thus, users are being restricted in their access, often times without their explicit knowledge and consent. This restriction in access results in not just filtering, but content blocking in pursuit of a goal, in this case, as stated by the actors mentioned before, the safety and well-being of the web user. While this action may be correctly referred to as censorship, the most commonly used wording by the UK government to describe their censorship is filtering [24].

C. Point of Filtering

Internet filtering can occur at many different levels, starting from the devices used to access the web up to state-level filtering where all content that enters or leaves a country through the high level internet infrastructure is monitored and filtered by the state or a comparable actor [25]. In case of the United Kingdom, web traffic is generally filtered at the Internet Service Provider (ISP) level, with each ISP being free in their choice of filtering mechanism. Filtering at this level in the web connection chain gives the controlling party the ability

to implement measures across all devices and lets them adapt quickly to changes in IPs, servers or tries of circumvention. Some schools in the UK use devices that monitor all traffic generated on their networks, giving them full overview of their students web activity [26]. This monitoring often extends to a child's home too through internet connected devices given to them by their respective school running a web monitoring software at all times, allowing for web monitoring at any point and any time [27].

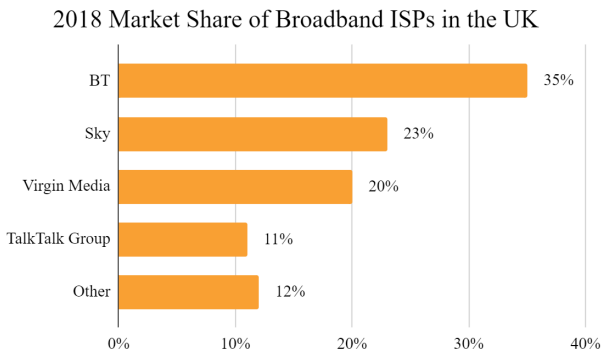

Figure 3. Market share of network providers in 2018 in the UK, based on the number of fixed broadband subscribers. Data from: https://www.statista.com/statistics/273412/market-share-of-uk-telecoms-operators-since-2007-by-fixed-broadband-subscribers/

D. Timeline of Internet Filtering

To get a deeper understanding of how the UK came to have such an exhaustive censorship infrastructure, one must look at the last 25 years of internet development in the UK. With multiple key events leading to the implementation of ever stricter censorship measures, the current state of web censorship is the culmination of a process that has still not finished, leaving room for a possible expansion in the future.

E. Early Beginnings

The arrival of the internet for the public in the beginning of the 1990s led to an adoption of the internet among a certain subset of technology enthusiasts [28]. Over the next few years, more and more private citizens started using the internet to communicate with each other and share information.

Among that trove of information some were being shared that broke the law by containing depictions of child abuse, leading to further investigation by the police. Working together with law enforcement in a multitude of countries, Operation Cathedral in 1998 led to the exposure of the biggest online child pornography ring known until then [29]. As a result of 105 concurrent warrant-based searches, over 750.000 depictions of online child abuse were found and confiscated, beating the previous record of 7000 pictures seized in one raid. Operation Cathedral led to widespread public awareness of the unchecked freedom the internet offered to everyone, allowing for the conduct of illegal activities with relative ease. Public opinion strongly favored stricter internet regulation and oversight, leading the government to set out in doing so [30]. In the year 2000, the UK government and the Department of Trade & Industry officially endorsed the Internet Watch Foundation, starting a close relationship with the charity and thus enabling it to become the internet watchdog it is today.

A major change in UK law was the introduction of the Sexual Offenses Act 2003, which specified older laws concerned with sexual offenses and amended several new ones regarding the relationship between the internet and sexual offenses [31]. Previous law set the maximum penalty for the sharing of child abuse images at three years; of those caught through Operation Cathedral, almost all were convicted under old law.

1) Mobile Internet Filtering: The first step to widespread web censorship came about in 2004, with the Independent Mobile Classification Body releasing its "IMCB Guide and Classification Framework for UK Mobile Operator Commercial Content Services" [32]. These guidelines roughly outlined the filtering policies most of the mobile network operators adhere to, setting the precedent for self-regulation and arbitrary censorship without a surrounding legal framework. The goal of the guidelines was to "offer a safe browsing experience for children" [33]. As a first step, prepaid handsets were targeted with these filter lists, with the only way of accessing

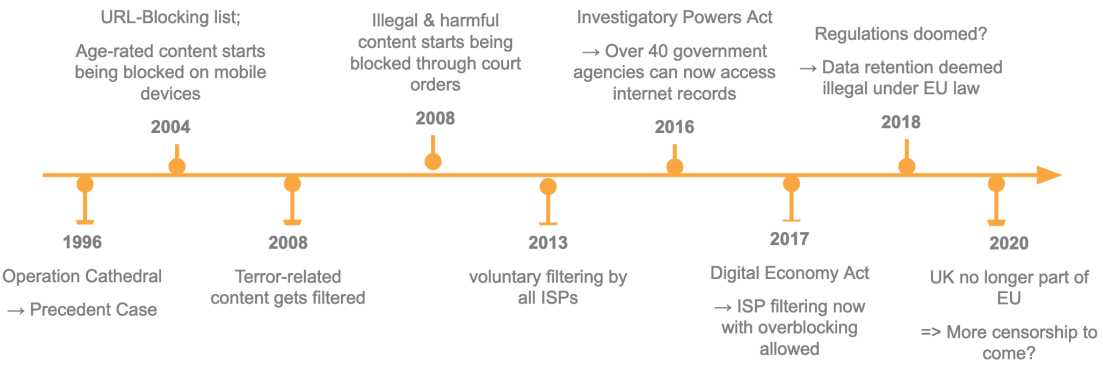

Figure 4. A timeline of web filtering.

the unfiltered web being age verification through personal identification.

Mobile internet filtering was expanded over the years to nowadays be an opt-out option on all mobile networks, forcing anyone who wants to access content not deemed to be suitable for under 18 year olds to register with their governmental identification. This default filtering is present even when the affected user is connected to a foreign network, for example when travelling through a country with unfiltered internet such as France or Germany.

2) Terror-related Filtering: 2007 saw two terrorist incidents affect the UK, with one failed car bombing in London and another failed one on Glasgow airport. The purveyors were found to be radical Islamist, of whom at least some radicalized themselves to some extent through the internet while also making use of the internet for bomb-building instructions [34] [35].

As a result of these terrorist incidents, in 2008 the UK government put into law the Counter-terrorism Act. While not directly concerned with web filtering in response to terrorist usage of the internet, it led to the expansion of the by then existing filter lists with pages related to extremist Islamic content. Until this day, websites deemed to be terror-related are being filtered, with the 2010-established Counter Terrorism Internet Referral Unit maintaining another list of sites that are supposed to "incite or glorify terrorist acts under Section 3 of the Terrorism Act 2006" [36].

3) Court Order Blocks: Under section 97a of the Copyright, Designs and Patents Act, Internet Service Providers in the UK can be legally forced to block access to web content if they "are found to facilitate internet copyright infringement (piracy)" [37]. Court order blocks require an extensive legal process in front of a High Court, making a rights claimant and monetary reserves necessary. The result is a legally binding block that has to be implemented by all ISPs in the UK, in the method of their liking. Penalties for non-blocking by an ISP do not exist. The court order blocks are among the few that are published online, making it possible to ascertain which sites are blocked and who made a claim against a particular site. Most websites on the list of court order blocks are blocked for copyright infringement reasons [38]. As of 2017, live blocking orders can also be sought in the courts, specifying the blocking of a live stream at certain times and on different hosts in order to prevent the public from streaming content such as pay-to-watch football games [39].

4) Default Filtering: Starting in 2013, almost all major ISPs started filtering web traffic generated in their networks. This came as a result of the acting government putting an increased focus on web security and protection of children from potentially harmful material [40]. With significant public support for web filtering, most major ISPs including BT, Talktalk and Sky moved to implement filter for all new customers in 2013, with existing customers being put under filter in 2014 [41]. The ISPs were not forced to implement content filters, but government and public pressure made them do so; legislation to make default filters mandatory was rejected in 2013 [42]. Stated goal was to implement web filtering for 95% of all households by the end of 2013 [43]. In a 2013 research by the BBC, filters used by major internet providers such as TalkTalk and Sky were found to block not only access to porn sites, but also sex education sites such as BishUK.com and Edinburgh's Women's Rape and Sexual Abuse Centre [44] [45]. A whitelist was proposed to solve the overblocking of acceptable and socially useful websites, but the current state of this whitelist is unknown. The extensiveness and unavoidability of internet filtering in the UK starting in 2013 led to the Reporters without borders considering them an enemy of the internet [46].

5) Investigatory Powers Act: The Investigatory Power Act, introduced in 2016 represents another step in the direction of complete web surveillance by the UK government. Implemented to more clearly define and regulate the surveillance powers available to the various entities of government, it also expands those powers in some ways [47]. Of particular interest is the ability to collect information in bulk by government agencies and the requirement for all Internet Service Providers to collect and retain their users' connection records for one year, while also giving a plethora of governmental departments insight into those records without the need for a warrant [48] [49]. Over 45 governmental departments and agencies are allowed access to internet connection records without a warrant, among them all sections of police, the Government Communications Headquarter and the Department of Health [50].

While the bill was passed into law in November 2016, the European Court of Justice ruled the aforementioned bulk collection and retention of user internet records unlawful in December of the same year [51]. As of today, the Act is in power with the data retention part not enacted upon. with the United Kingdom in the process of leaving the EU, the future regarding bulk data collection and retention in the UK is unclear.

6) Digital Economy Act: The biggest legal change to the UK webspace came about with the introduction of the Digital Economy Act (DEA) in 2017. In the context of this paper, it introduced two major changes affecting the openness of the internet: Age verification and default web filtering [52]. Under the Act, commercially operating websites offering pornographic content would have to ensure any user accessing their site is over the age of 18, lest they pay a fine. Any noncomplying websites would be put on a block list. Through multiple technical difficulties and the BBFC's inability to rate every pornographic website on the internet, the age verification scheme was dropped end of 2019 in order to be replaced with a wider regulatory scheme in the future [53].

With overblocking present in most existing blocking mechanisms up until the introduction of the Digital Economy Act, the DEA makes it legal for ISPs to overblock websites in order to block other ones. Article 23 (3) explicitly states: "The steps that may be specified or arrangements that may be put in place under subsection (2)(c) include steps or arrangements that will or may also have the effect of preventing persons in the United Kingdom from being able to access material other than the

offending material using the service provided by the internet service provider." [52]. Subsection (2)(c) gives ISPs the right to choose the method of blocking, while section 3 guarantees them no legal repercussion if their blocking leads to over sites being overblocked. As of today, the only way to appeal an overblock is by requesting an unblocking at the ISP directly, with an unblocking not guaranteed. Additionally, siteowners are not notified if their site is being blocked, making self-discovery as a result of missing traffic the only way to detect if one is being blocked by one or more ISPs.

With web filtering already implemented across nearly all ISPs in 2017, the only step missing was the legal requirement to do so. The Digital Economy Act forces all ISPs to blick websites containing adult content by default, giving internet users only an opt-out option to disable filtering.

F. Illegal vs Harmful Content

An important topic of concern is the UK government's intent to block access to illegal as well as harmful content. As seen above, the monitoring and restrictions were put into place to prevent the sharing and consumption of illegal material, in most cases child abuse content. However, internet filtering as of today is more often concerned with filtering material deemed harmful for a specific subset of people.

Therefore, a distinction must be made between what is filtered, and in his 2001 research, Yaman Akdeniz does so by specifying illegal content to be one "criminalized by national laws, while [harmful content] is considered as offensive or disgusting by some people but certainly not criminalized by national laws" [54]. This content might be certain acts deemed sexually indecent, certain opinions or sexual orientations.

The British Board for Film Classification (BBFC), which since 2013 has been responsible for the classification framework mobile network operators use to self-regulate commercial website content accessed through their networks, uses a binary classification system to determine whether content should be accessible by children, defined as any person below the age of 18 [55]. Among the types of content considered suitable for adults only the BBFC lists drugs, sex and violence. However, the mere "Repeated / aggressive use of [the word]'cunt'" has been deemed unsuitable for minors and any website seen to fall under this classification will therefore be blocked on all mobile networks [56].

G. Internet Watch Foundation

The Internet Watch Foundation (IWF) is a key player in the regulation and censorship of the UK internet. Founded in 1996 as a result of police interest in the dealings of Usenet newsgroups suspected of sharing child pornography, the IWF is a joint effort between various stakeholders such as the ISPs, the Metropolitan Police and charity bodies in identifying and "combating criminal material on the Net" [57].

24 years later, its mission in finding and preventing access to depictions of child abuse on the internet is still ongoing, with the IWF providing one of the major URL lists, according to which websites are being blocked in the UK [58]. Although automated technology might be used too, a staff of human analysts is responsible for identifying URLs on which child-abuse content is hosted and then adding them to the URL blacklist. Being estimated to contain just 900 URLs in 2009, the URL list is said to contain over 108.000 unique URLs in 2018 [59].

A 2009 study by Richard Clayton estimated that over 25% of pages contained in the URL blacklist were file hosting sites that could very well be used for non-illegal services [60].

H. UK Webspace after Brexit

It is therefore important to measure the extent to which the internet is monitored now and be aware of any changes in data protection laws that might come as a result of the United Kingdom leaving the European Union.

IV. Levels of Filtering

Generally, there are multiple levels of filtering present in the UK. The levels of filtering in this case refer to the types of content being blocked and the parties interested in blocking access, relying on their legal bases to do so.

The most encompassing filters are those intended to block out illegal content such as child pornography. They are ever-present and often used in combination with take-down requests for websites depicting such content. Filters for such illegal content were among the first to be implemented and find broad public support.

The second level of filtering can be considered that of the Public Court Orders. Dealing mostly with copyright content violations, they make certain web content illegal as a result of an judiciary process and ruling. Implementation is legally binding.

On the third level of filtering are ISP filters, used to prevent the spread of harmful content to children as well as the distribution of certain ideologies and thinking. ISP filters are based on filter lists as described above, with potential for subjectivity in the blocking.

Mobile filters can be viewed to be on the same level as the ISPs, as they also implement ISP blocking, often by the same providers, but in the case of the mobile filters, with the premise of a general filtering of all content unsuitable for minors.

Filtering in the UK is also present in Local Area Networks, with school and public computers in libraries subject to even more scrutinized internet access, often allowing only whitelisted content to be viewed. This represents the fourth level of filters, which can often not be circumvented easily and targets subgroups specifically.

V. Web Filtering Mechanisms

Web filtering can be based on a plethora of different filtering mechanisms, of which the most important ones will be introduced and discussed in this section. The exact implementations used by each internet provider are not known, but can in some cases be assumed based on their HTTP response. Often, third party companies provide the technical know-how and manage the filtering in place of the ISP, absolving the ISP

of personal responsibility in case of overblocking and letting non-governmental actors inspect UK citizens' internet traffic [61].

A short overview of the most commonly used blocking techniques:

1) IP-based blocking
2) URL-based blocking
3) Deep Packet Inspection
4) DNS-based blocking

All of the blocking methods mentioned above can be circumvented in one way or another, with the use of proxies or a Virtual Private Network being among the most common [62]. The prevention of blocking circumvention is part of the analysis below.

A. IP- and Protocol-based Blocking

IP-based blocking prevents an user from accessing a certain IP address. This requires the blocking party to have control over the user's internet connection and being able to identify what IPs they are trying to reach and then either redirecting or returning an error when that user tries to access content from that specific server [63]. TCP blocking restricts access to certain TCP-port numbers that are associated with a service or server to be blocked [64]. IP-based blocking does not distinguish the content on a server, but restricts access to it completely, thus making it a very broad blocking method with potential for great overblocking. Additionally, IP blocking is often associated with the blocking of a range of IPs in order to fully prevent access to a specific server, thereby also blocking access to servers that should not have been blocked but got caught in the range of IPs.

This type of blocking is easily circumventable by the administrator of the blocked server through either a change of the IP address associated with the server or using content delivery networks with constantly changing IPs.

B. URL-based Blocking

URL-based blocking uses the web URL requested by an user to determine whether access should be granted or prevented. It compares the URL included in the HTTP request to given lists of URLs to block and can then either let the HTTP request through, block it and return an error or send the user to another website instead [65]. Doing so requires the capability to intercept traffic and check HTTP request quickly in order to guarantee internet usability, making significant computational power necessary [66].

URL blocking has its limits where the URL associated with a server changes often and where deep links shall be blocked, while high-level links shall stay available. An example of this could be the blocking of certain content on a website such as www.example.com, which should generally stay available while blocking certain deeper links (e.g. www.example.com/indecentphotos) containing material deemed to be improper.

C. Deep Packet Inspection

Deep Packet Inspection (DPI) is one of the most invasive methods of internet filtering as it involves the inspection of all data packages instead of only the data header sent by an user; this happens at a point in the connection between the user and the server they are trying to reach [67]. DPI allows for real-time identification and targeting of content and users alike, allowing for extensive profiling and the possibility of user-specific censorship. As such, DPI is very resource intensive, with a plethora of current research focusing on efficiency gains in regard to the internet infrastructure level inspection of packages [68].

Deep packet inspection is one of the tools used by internet providers to determine what content an user is trying to access and allocating them different bandwidths if they are, for example, trying to access a music streaming service that should not account towards monthly data caps according to the individual's data plan [69]. Through Unique Resource Identifiers, it is possible to detect particular resources being requested in an users internet traffic.

In the European Union, DPI is the basis for the widespread violation of net neutrality by ISPs and is therefore illegal [70]. With the United Kingdom still a member of the EU, those laws still apply, but are not being enforced.

D. DNS-based Blocking

DNS-based blocking works by controlling and modifying DNS queries. In that case, the DNS resolver is used to perform DNS look-ups while at the same time checking if a requested server name is on a block list. If it is, instead of returning the correct IP of the server the user is trying to access through a requested URL, the user can either be returned a DNS error or be redirected to an entirely different server [71]. One example of such blocking is suspected to have been the case in the blocking of the website of the German hacker association "Chaos Computer Club" in 2014, which for users of more than three ISPs was unreachable through its URL, but still reachable by directly using the IP associated with it [72].

E. Third Party Blocking Solutions

For the physical implementation and oversight of the blocking found on their networks, many British ISPs make use of third party companies that offer products capable of real-time-filtering for every user in an ISP's network. ISPs generally do not disclose what filtering methods they make use of and by whom they are implemented, however, some have been identified. To note here are primarily the solutions by Symantec, an US-American software company with a focus on cybersecurity, nowadays owned by Broadcom and rebranded as NortonLifeLock [73]. With their Rulespace software, British Telecom, EE and Telefonica were able to implement web filtering in their networks [61]. Rulespace has been criticised for overblocking from the beginning [74] [75].

The ISP TalkTalk uses a blocking software by the Chinese technology company Huawei called Homesafe. Huawei has been criticized multiple times for alleged ties to the Chinese

government [76]. It is feared that with Huawei in control of the blacklists and blocking, political content could also be targeted [77].

VI. Measuring Internet Censorship

With the internet censorship sphere in the UK split up in the ways described above, the research done here provides some valuable insight into the actual blocking being done. No unified governmental oversight and no centrally organised publicly available option of determining whether a website is being blocked exist. Therefore, non-governmental entities have started compiling data on censorship in the UK, with the Open Rights Group (ORG) and the Open Observatory of Network Interference (OONI) being at the forefront of data collection. Firstly, a statistical oversight of the blocking actions will be presented, followed by a deeper dissection of a subset of blocked sites. One of the biggest paradigms fueling this research is overblocking, which is apparent in the filters of various ISPs. This eventually leads to the aforementioned restriction of personal freedom in pursuit of perceived harm reduction.

A. ORG Blocked Project

The Blocked project is an initiative from the Open Rights Group (ORG), which aims to evaluate the status of website blocking in the UK. Established in 2014, in response to the missing possibility of actually checking whether your website is being blocked or not, it is one of the only tools available to the public that can test for blocks on all major ISPs. The idea of the project is that owners of a webpage should be able to test whether their page might be unreasonably blocked. In case that the owner of a webpage sees that his page is blocked, the project offers forms, which can be used to contact the according ISPs to ask for an unblocking. Additionally, this feature also enables internet users to check whether some webpage is blocked by any of the ISPs in the UK. This also allows to get some understanding about the procedures of blocking mechanisms and techniques, which are applied by the ISPs.

To check whether a provided webpage is blocked, the project has set up a number of probes on different ISPs in the UK. In this context a probe is a computer which needs to be constantly connected to the internet and which runs the ORGs software for testing the blocking status of a website. Through its website, one can insert an URL and the probes try to access the page [78]. The available probes will then run a test to check whether the requested webpage is accessible. For webpages that have been checked before, previous results can also be seen, letting the user know whether an ISP did not block their site before or had their site blocked. The projects webpage also reports statistics on the number of blocked pages per ISPs and successful cases where pages have been unblocked by the ISP on request.

A Blocked.org probe is a Raspberry Pi computer, on which the backend software from the ORG is installed. A probe can then measure the availability of webpages on the network it is connected to. Volunteers are also able to buy or setup their own probes [79]. The ORG has setup probes on all major ISPs in the UK and also configured them with different filtering preferences. For example British Telecom (BT) and some other ISPs provide a filtered internet access as well as an unfiltered service (Section VIII-A).

A blocked page is determined via a set of specific rules for each ISP. The rules are manually adjusted by the ORG collaborators [80]. It is checked whether a page is redirected to some specific blocking URL, or if the HTML body represents a blocking page. As the blocking rules for each ISP are added manually, it is possible that the rules for the current blocking mechanisms of an ISP are not amended and a block might therefore go undetected. For the Andrews&Arnold Internet Service Provider's (AAISP) network no blocking rules are specified, leading to no pages marked as blocked. AAISP is also the single major ISP that openly made free and unfiltered internet access their mission, offering no filtering or logging at all. For that, they have been criticized by the UK government. [81] However, as the HTML file from the response is checked for a redirection to a specific blocking page, falsely reported blocks should not occur. Those false positive blocks cannot be determined without direct access to the ISP. Therefore, the testing methodology used here could not account for these cases. Additionally, in case of an unsuccessful page request, the error type was reported in the status. The full list of ISPs, on which measuring probes are setup is displayed in Table XIV.

It is possible that an ISP does not redirect the user to some customized blocking page, but instead blocks the page by causing some network error. The Blocked project reports an error type in case the investigated page did not load correctly. An error can, however, also occur in case some failure occurs on the server of the requested site. OONI uses a control probe, which simultaneously to the testing probe, loads the same websites from an uncensored network, in order to approximate whether the investigated website is actually available. In case of the Blocked project it is uncertain whether the reported website loading errors might indicate an undetected block, as no control probe is available.

As the service of the Blocked project is free to use an open to anyone, until January 2020 around 6 million webpages have been tested. However, for this reason the tested pages are not selected based on some specific criteria, but only represent the interest of the internet users which know and use the service. Therefore, the distribution of the tested pages is not representative of actual web traffic in the UK.

In Section VII-A we describe our methodology of retrieving data of the project for our analysis.

B. OONI Internet Censorship Probes

The Open Observatory of Network Interference (OONI) is a project run by a like-minded community intent on detecting censorship and internet interference. The goal of the project is to provide a way to test whether a network shows any signs of internet interference. To do so, the project has designed a

number of specific tests which measure the accessibility of webpages or the performance of a network. [82]

Similarly to the ORG Blocked project, OONIs internet measurements also rely on probes, which run a number of tests. The goal however, is to measure network interference across the world and not only at some predefined physical location. Therefore, probes are not primarily setup by the core team, but the project relies on volunteers, who run probes. This can be done by installing the provided open source software on some device which has access to the internet. The OONI software is available for a number of platforms, including an app for Android and iOS devices. The incentive for volunteering is the ability to test one's own network and determine whether one might be the target of some network interference actions. Volunteers can manually adjust which tests should be performed by their device. With the consent of a volunteer, test are performed regularly, which allows the tracking of possible changes in interference actions in some network. All data from the performed measurements is afterwards send to the OONI servers and publicly available. However, the data does not disclose concrete information about the devices that ran them. Still, for volunteers in some countries with oppressive regimes it might be a risk to run such a probe, as trying to access blocked content can be enough to trigger repercussions for the volunteers.

The majority of measurements were performed by probes from western countries. However, some measurements data is available for most countries in the world.

The project provides an online explorer tool, which allows one to navigate through measurements by country, or to search for specific measurement reports.

In total the OONI software has 17 tests, which aim to answer the following questions: [83]

- Which websites are blocked?
- Which instant messaging apps are blocked?
- Is Tor blocked?
- Are VPNs blocked?
- What is the speed and performance of my network?
- Are middleboxes present on my network?

For our investigation we are interested in web censorship in the UK. One test which measures this is the so called Web Connectivity Test. It aims to detect whether a website is blocked for the testing probe. To determine the blocking status of a site, the site is first accessed from an uncensored vantage point. The connection information, including the IP address of the site, the headers and the length of the returned response bodies, are then sent to the probe. Simultaneously, the probe also tries to establish a connection to the website. In the final step the connection results from the expectantly uncensored control probe and the testing probe, in the investigated network, are compared for inconsistencies. In case of inconsistencies, a website will be reported as potentially blocked. Anomalies are checked on the TCP/IP, DNS and HTTP connection stages. Therefore, in case an inconsistency is detected, the connection stage on which the deviation occurred is reported as the suspected stage of censoring. Unfortunately, the calculated anomaly attribute measurement report may also contain false positive results. If for example, the control probe happens to fail to resolve an URL, this can then lead to an inconsistency in the testing and controlling probe test responses and is therefore reported as a block. [84] As the full connection details are reported for each measurement, it is possible to calculate the blocking status using some customized rules. In Section VII-C we describe the rules for website block determination.

In contrast to the ORGs Blocked project, the tested websites are from a manually curated list of potentially censored websites from the Citizen Lab [85]. The user can also insert some URL in the OONI application which should be tested, however, the site will only be tested by his probe. It is possible to propose new websites to the list, by filling out a form on the project page. Therefore, the list of tested websites is by far not as extensive as in the ORGs Blocked project. As the websites from the testing list are meant to be tested by probes all around the world on a regular basis, it would not be feasible to enlarge the list of tested pages extensively.

The full description of the Web Connectivity Test can be found on in the GitHub repository of the OONI project [84].

VII. Methodology

In order to generate valid data, multiple sources have been considered and made use of. This leads to varied measurement data, allowing for greater data exploration and analysis.

A. Loading Measurements Data from Blocked ORG

The blocking status for each tested website can be downloaded as a compressed CSV file form the ORG Blocked project website [78]. The downloaded file is created directly from the backend measurements database of the project. However, it only contains the status of the latest measurement for each tested URL on each network and not all performed measurements.

We have downloaded the latest URL statuses for each tested website, as of the 17 January, 2020. The first measurements are from May 2014 and in total 80 million measurements for 6.7 million URLs are reported. In total, measurements are available for 23 distinct ISP and configuration settings.

The extracted CSV file has a size of 8.1 GB and was too large to load it into the main memory of our local machines. However, as the dataset is structured we decided to load it into a database system (DBS). As we wanted to work collectively with the data, we decided to use a remote DBS instead of setting up a local DBS. After comparing existing cloud DBS services we decided to use Google's BigQuery. One of the advantages was that it wasn't necessary to decide on one server type, on which the DBS should run, upfront. Instead Google automatically allocates the required computational resources when querying. This aims to deliver fast and scalable query executions. Upon sign-up the user receives a 12 months free trial, including a budget of 300$ for service costs. Storing data is initially free of charge up to some amount. Google charges service costs for the amount of queried data. The runtime of

the executed queries are of no importance for the costs. The costs for querying 1 GB of data is only marginal with around 0.5 US $ cents. For our research the provided free credits were sufficient and we did not face any additional charges.

The downloaded CSV file is in a structured form and did not require any upfront data cleaning. Creating tables from structured CSV files is straightforward in BigQuery. However, table can only be created from uploaded CSV files up to 10 MB. Therefore the downloaded file exceeded this limit. To be able to load the data into a table we first had to upload it into the Google Cloud Store. This is an additional data storage service, but the amount we needed was also free of charge. From here it was possible to create a new table from the uploaded CSV file in BigQuery.

When creating tables it is also possible to partition them by some ordinal variable. We partitioned the table by the date of the measurement. Partitioning a table allows to restrict the queried dataset to only one day of measurements and the costs for the query are much lower, compared to querying the full 8 GB of data. Therefore, when experimenting with a query we first only executed it on one day of measurements, by specifying the date in the SQL *WHERE* statement, and only afterwards let the final query run on the full dataset. Running queries in BigQuery was very fast and just rarely took longer than a couple of seconds. We either saved the results as a new table for further analyses or downloaded the results as a CSV file.

Due to the public URL checking service of the Blocked project the available dataset includes a very large number of tested URLs. However, as described, only the latest blocking status for each URL and ISP is reported. Therefore, it is not possible to compare the blocking status over time. One possibility would be to pull the dataset multiple times during some predefined time frame and check for changes of the blocking status for some URLs.

The blocking status of the measured sites also depends on the manual updating of the blocking rules for each ISP. We were not able to estimate how accurate those rules are and whether all blocking mechanisms are detected. Also it was unclear whether a reported page loading error for some ISP might also indicate a potential block. Therefore, we were only able to limit our analysis to the reported blocks in the dataset (*status = blocked*).

B. Fetching Measurement Data from OONI

OONI provides a free API (Application Programming Interface) in order to retrieve the measurements reports of all tests which were run by probes [86]. We used this API to pull data from the performed Web Connectivity Tests[2].

To retrieve the measurements data, we have used two types of API calls. The first call retrieves measurements data from a meta database (https://api.ooni.io/api/v1/measurements). This API call accepts further parameters, which allows to specify the measurements of interest that should be returned. First we specified to retrieve measurements only from the Web Connectivity Test (*test_name=web_connectivity*) and set the parameter *probe_cc*, which indicates the location of the measuring probe, to only include measurements from the UK[3]. To avoid running into server timeouts, we have called the API only for one day of measurements data and later combined the daily measurements into one datasets.

Figure 5 depicts the API response for an example call of the OONI measurements metadata. The *results* field is a list, which contains the metadata for the returned measurements report. The metadata already includes the *anomaly* attribute, which indicates whether the tested site might be blocked, however, as described in Section VII-C this information was not sufficient for our analysis, due to the high number of false positive results.


```
{
  "metadata": {
    "count": 30,
    "query_time": 7.184073448181152
  },
  "results": [
    {
      "anomaly": false,
      "confirmed": false,
      "failure": false,
      "input": "http://www.eccouncil.org/",
      "measurement_id": "temp-id-370876357",
      "measurement_start_time": "2020-01-01T08:23:12Z",
      "measurement_url":
"https://api.ooni.io/api/v1/measurement/temp-id-370876357",
      "probe_asn": "AS5607",
      "probe_cc": "GB",
      "test_name": "web_connectivity"
    },

   ...

  ]
}
```


Figure 5. Exemplified extract of the JSON response for a list of measurements from the OONI API.

The metadata, however, also includes the URL to the full measurement report. We iterate through all entries in the *results* list and then perform a further API call to the measurement report URL. The response is again a JSON file, which contains all the data from the Web Connectivity Test. From the full measurements report we extracted both the connection attributes from the control probe and the testing probe. For the control probe we have extracted the HTTP status code from the measurement, the page title of the returned website and three aggregated attributes, which indicate whether a failure occurred during some connection stage. For the testing probe

[2]The full code which we used to retrieve the data from the OONI API, the analysis of the data and the code for the further clustering of the web content can be found on GitHub: https://github.com/jkrol21/UK-Internet-Censorship-Research

[3]The OONI country parameter for the United Kingdom is depicted as *GB* and not *UK*. Therefore, it would indicate that it would only include England, Wales and Scotland, while the UK would also include Northern Ireland. As OONI did not report a separate country code for Northern Ireland and in the OONI Explorer the abbreviation GB is described as United Kingdom, we assume that probes from Northern Ireland are also included in our retrieved measurements.

we have extracted information on the platform of the probe, its network type, the IP of the used DNS server and also the attributes indication whether some failure occurred during some connection stage. Finally, we also fetched the results of the consistency checks from the control and test probe and the resulting blocking attribute, which indicates on which level the possible blocking occurred. The consistency attributes are determined using the logic in OONIs's Web Connectivity Test. It is also possible to derive different rules for the detection of blocks, as the full connection attributes are provided. For our analysis we have decided to use the existing logic for determining blocks, but cleaned the dataset in order to limit the amount of false positive blocks.

The measurement report also contains information on the Autonomous System Number (ASN) of the probe. The ASN represents an identification number, which is assigned to a set of IP addresses, which are mostly associated with one network. We use the data from GGPview [87] in order to estimate the ISP behind a given probe's ASN.

Additionally, we used the IP address from the DNS of the testing probe to further estimate the possible network. To resolve this information we used the data from dbip [88] to get the network information for each DNS IP address.

Unfortunately, the API restricts the number of accesses from one IP address and prohibits the inspection of a wider timeframe through bandwidth restrictions. In our case, in one hour we were able to retrieve around 3,000 measurements. Therefore, we were not able to retrieve all performed measurements in the UK, but had to limit the timeframe for our analysis. We have collected a dataset of 107,546 measurement reports for the time period from 1 January, 2020 until 31 January, 2020, from probes in the UK.

Table I depicts the ISPs, which aren't marked as mobile networks in the OONI measurements and Table II depicts the mobile ISPs from the OONI measurement.

C. Cleaning of the OONI Measurements Data

The *anomaly* attribute from the OONI measurements report reports a suspected block in many cases, although when looking at the HTTP response of the testing probe, it becomes apparent that the requested page did actually load correctly. To limit the amount of false detected blocks we have excluded some measurements, which contain some errors and are likely to be falsely classified as blocks

One problem in the calculation of the anomaly attribute is the measurement from the control probe. In order to determine whether an interference occurred in the tested network, the results from requested pages of the tested probe and the control probe are compared. It is determined whether the DNS response, the TCP/IP connection or the HTTP(S) response deviate between the tested probe and the control. In case an inconsistency is determined, the website is marked as potentially blocked (anomaly = True). However, it also occurs that the measurements from the control probe fail. For example, in our collected measurement dataset, the *YAHOO-Mail* page has been marked several times as potentially blocked. However, the HTML responses show that the regular YAHOO page has been returned to the tested probe and the error occurred in the control probe [4].

In the measurements report, the HTML page title from the control probe website connection is reported. When looking at some of the detected anomalies, we saw that for various tested websites the page title from the control probe was *Attention Required — Cloudflare*. This indicates the control probe ran into some connection error. This then causes that the HTTP response of the control probe is not going to match the result from the testing probe. As for the detection of an actually censored website, it is necessary that the investigated page is actually available, we have excluded all cases where the page title of the control probe depicts the Cloudflare error page text.

Another example for a false positive determination of a blocked website is a different content which is displayed for users from different regions. For example the page *www.pandora.com/* (a music streaming service from the United States) is sometimes also determined as blocked in our dataset. When we accessed the page from our local ISPs in Germany, the page displays an information that its content is currently only available for users in the United States. Some of the OONI measurements reported this page as blocked. However, measurements from the same ASN network and using the same DNS have report the pages as blocked in some experiments and in some as not blocked. The blocking reason is each time a difference of the length of the HTML page from the HTTP response. As the length of the returned page matched in some cases for the testing probe and the control probe, we assume that OONI uses control probes in different locations. In case *www.pandora.com/* is reported as blocked, the control probe is probably located in the United States and the probe retrieves the full content of the page. When the page is reported as not blocked, the control probe is probably also located outside the United States and therefore the retrieved content matches the page when accessed from the United Kingdom. We have manually detected this anomaly in the reported block pages and removed the measurements of *www.pandora.com/* from the blocked pages in our dataset. It is possible that more false positive cases of similar types are included in our dataset.

Before cleaning, the dataset contained 107,546 measurements and 6,300 detected blocks. After we applied our cleaning rules the dataset contained 100,343 measurements and 5,364 detected blocks. We use this cleaned dataset for our further analyses.

Not all the websites which have been marked are not accessible to the internet user. The ISPs have also implemented age verification pages. Figure 6 shows one example page of such an age verification page. The user is required to enter his credit card details in order to prove that he is over 18 years and to be allowed to access the requested page. In the case of the O2 network, the user can also show an official ID in one

[4]One example for a measurement with a reported block due to a failure of the control probe: https://api.ooni.io/api/v1/measurement/temp-id-371928788

of stores from the provider. Once verified, according to O2, the user will be able to access age restricted pages without a further verification [89].

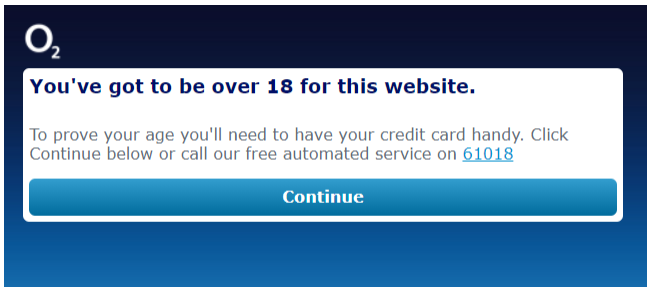

Figure 6. Example age verification page from the O2 mobile network. Taken from the HTTP response of one example OONI measurement: https://api.ooni.io/api/v1/measurement/temp-id-377435600

In the OONI measurements those sites would also be counted as blocked, as the returned content differs from the loaded page by the control probe. In the case of the measurements from the Blocked project, it is unclear to us, whether the blocking rule for the ISPs also classify an age verification page as blocked.

There is no clear way to determining blocks in from the OONI measurement reports. With our approximation we aimed to filter out some of the apparent false positive blocks. However, it still not definitely clear how many false positive blocks are in our approximation. The case of false negative blocks is less likely from the OONI data. However, it might also be possible that the seemingly non restricted control probe would be redirected to the same blocking site as the testing probe. We omitted this case, as it seemed very unlikely to us, but if it occurred that our number of detected blocks would also be over estimated.

It is also possible to export the all OONI measurements into a database from the OONI project [90]. The full dataset is, however, around 7 TB of size. This approach was outside the scope of our work and we, therefore, continued to use the restricted OONI API.

D. Scraping of Websites in the Blocked Dataset

For further analysis, the URLs of all blocked websites on the Sky network were collected and aggregated by category.

In order to make sense of the thousands of websites determined to be blocked, an analysis focused on commonalities and differences was deemed appropriate. The goal was to not only find overblocked websites, which would be additional proof of overblocking but to find similarities among these websites in order to find out why they were being blocked and how overblocking could have been prevented; assuming motivation to do so on behalf of the ISP.

Web scraping was determined to be the most useful task at hand, with the goal of collecting as much text data from a given website as possible. Web scraping, often also known as data mining, is the process of querying a server, searching through the HTTP response of that server for information one wants to gather and parsing that to a file [91]. Web scraping enables one to gather information that cannot be accessed through an API, and is often not immediately present. With many of the websites scraped in this research being of completely different makeup and build. Therefore, web scraping by taking the content one would normally see on a screen makes the most sense, giving us exactly the content that would be displayed when viewing that particular site.

With many of the sites being very different in their content and build comes the challenge of building a web scraper capable of extracting information from a variety of sources. The base language of the webscraper created for this research is Python, for its multiprocessing scalability and the packages offered for it [91]. BeautifulSoup, a package for Python 3.0, offers the ability to build a parse tree from accessed websites, making use of the HTML data in order to generate scrapable content. It enables the user to make sense of the HTML content in a structured way and extract text, logic or other information from it.

In the case of our research, we tried to access a website's description, keywords, title, tags, heading, content and overall text. A website's description consists of all content in meta tags or attributes called "description", with the BeautifulSoup Style document as the base file to extract data from. Similarly, keywords were extracted through the content attributes. The description often contained full sentence structures, while keywords indicated a website's contents in words with similar content, often to improve that website's search indexing. A website's title was also scraped, giving information on a website's main focus and validating the contents of the rest of the scraped text.

Headlines and headings were scraped by going through header tags and accumulating all text found within them. Finally, to find all texts of websites with irregular structure, all text from p tags was scraped and all text from that page with the exclusion of a list of tags parsed as well, guaranteeing textual info from all websites taking into account the vast differences in HTML tree structure.

Scraping a website could only be achieved by making sure the URL requested actually returned a valid HTTP response, which could be checked by using the status codes returned. Only in case of a returned status code was the request successful, with status code 200 returning workable HTML content. The web scraper was written to be executed in multiple processes simultaneously, allowing for great timesavings. The code was run either locally, with all internet requests going through an encrypted VPN connection or on a Google Server through Google Colaboratory. Due to physical processor limitations, running on Google Servers was more advantageous, allowing for a higher number of parallel processes due to more CPU cores being available.

Scraped data was collected in a dataframe and exported as a CSV.

E. High Court Order Blocks

The *Blocked.org* measurements data also allowed us to collect information on the status of websites ordered to be blocked by an UK High Court. This was done over all ISPs, with accessibility and HTTP response being recorded. Errors were also saved, in case accessing a website on a particular network returned an error-code.

VIII. Analysis of Measurements Data

A. Blockage per ISPs

As the blocking of websites needs to be implemented by each ISP, it is of interest, to which extent ISPs censor websites and how the amount of blocked sites varies between them.

The OONI measurements report contains an attribute, which indicates whether the tested network is a mobile network. In total 5.7% of the measurements in our dataset were on mobile networks. To compare the blockage between mobile and stationary network providers we calculated the blockage of tested website per ISP. In Table I we show the non-mobile networks and the share of distinct URLs which were detected as blocks and in Table II we show the same results for mobile networks. The share of blocked URLs varies considerably per ISP. However, in the measurements the URLs that were tested by probes on each network vary. Therefore, it is uncertain, whether the share of blocked URLs per ISP might be diluted by the selection of the tested websites. For example probes on one ISP might have tested websites which depicted adult content and were therefore more likely to be blocked, while probes on another network might have tested none controversial content. During a measurement the probe performing the OONI Web Connectivity Test only tests some fraction of the possible URLs from the Citizen Labs curated list of URLs. Unfortunately, the amount of tested URLs per ISPs was too small in order to only look at the intersections of tested URLs across the ISPs.

The list of tested ISPs also varies, when compared with the probes setup by the Blocked project (Figure 7). Daisy Communications for example is a provider of network services for businesses and Jisc Services provides network services for educational institutions [92], [93].

The amount of tested URLs in the Blocked projects measurements is considerably higher. Also in the Blocked project a submitted URL is tested by all available probes, which are mostly setup on the most popular networks for private customers.

The larger amount of tested URLs enabled us to compare the restrictiveness of the ISPs in a larger extent. For an uniform comparison of the ISPs, we only look at URLs which have been tested by all ISPs. As the probes for each network configuration have not tested the total amount of URLs in the dataset, we only limit our analysis to the 15 ISPs with the most tested URLs. From those we excluded the provider

Table I
Top 10 non mobile ISPs from the OONI Measurements with the most measurements

ASN Network	Measurements	Tested URLs	Blocked URLs	% URLs blocked
Daisy Communications Ltd	38,496	1,222	50	4.09%
Jisc Services Ltd	17,572	1,638	81	4.95%
KubeNET	6,241	1,579	91	5.76%
BTnet UK Regional network (BT)	4,902	1,093	63	5.76%
Virgin Media	3,326	997	51	5.12%
M247 Ltd	3,221	1,556	57	3.66%
UK Internet Service Provider (BT)	2,429	1,147	43	3.75%
Sky UK Ltd	1,879	504	54	10.71%
ONLINE S.A.S.	1,643	610	238	39.02%
TalkTalk Communications Ltd	1,223	622	60	9.65%

Table II
Top 7 mobile ISPs from the OONI Measurements with the most measurements

ASN Network	Measurements	Tested URLs	Blocked URLs	% URLs blocked
Hutchison 3G UK Ltd (3)	1,389	758	49	6.46%
Telefonica UK Ltd (02)	1,271	341	26	7.62%
EE Ltd	1,135	533	8	1.50%
Vodafone Ltd	722	488	27	5.53%
Virgin Media	235	99	0	0.00%
OVH SAS	196	75	1	1.33%
EQUINIX BRASIL	192	53	3	5.66%

AAISP, as no blocking rule is specified for the service and therefore possible blocks could not be detected (see Table XIV. We then calculated the set of blocked URLs, as all the URLs which have been marked as blocked by at least of one of the investigated networks.

This resulted in a list 223,000 potentially blocked URLs. Using those URLs we compare the restrictiveness of the different ISPs and their network configurations.

As depicted in Figure 7 BT with their strict network configuration blocked over 80% of the tested URLs, which were blocked by any of the other networks.

BT was also the only network provider in the dataset with a stricter than regular censoring service. This package is designed as a parental control service and allows to set categories for content that should be blocked [94].

As to the standard setups from the other network provider, their restrictiveness is mostly in a similar range. Only the restrictiveness of Plusnet is well below the other network service providers. What stands out the very low restrictiveness of the unfiltered packages of the network providers. Compared to the standard configurations, only a handful of websites is blocked.

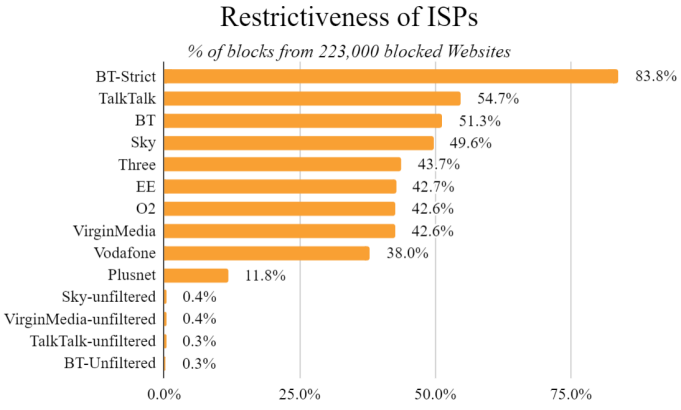

Figure 7. Comparison of determined blocked sites per ISP from the measurements from the Blocked project. Only considering the 223,000 websites which were blocked by at least one ISP and which were tested by all depicted ISPs.

B. What Blocking Mechanisms are Used?

The Web Connectivity Test from OONI approximates the used blocking technique, by comparing the controls and testing probes connection results at different connection levels. The lowest layer on which a deviation between the controls and testing probes results occur is then returned as the likely level of filtering.

In Figure 8 we show the distribution of blocking techniques for the detected blocks. As depicted in the figure, the most blocks occur on the highest connection stages (HTTP). Blocks through manipulating the DNS responses are applied second most by the ISPs in the measurements. IP addresses of the hosting servers seem to be only very rarely blocked.

To further approximate the blocking mechanisms of the ISPs we compare the blocking of websites which used the unencrypted HTTP vs. blocks from websites which use the encrypted HTTPS protocol.

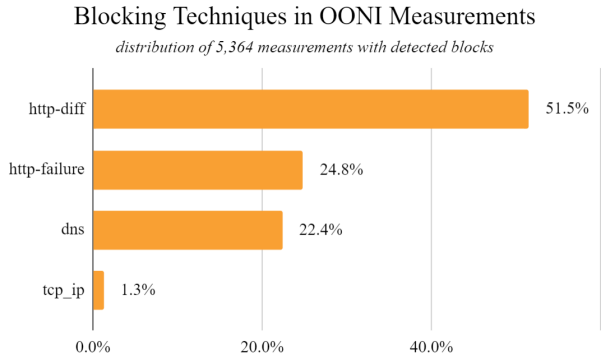

Figure 8. Distribution of blocking techniques in the detected blocks from the OONI Measurements.

The share of HTTPS URLs in the Blocked project's measurements is less than 1%. This is probably due to the reason that when an URL is inserted into the URL submission field of the Blocked project website without the prefix 'https://' it is by default called using the unencrypted HTTP protocol. On the project site, it is also described that in case a HTTPS site is blocked their system would most likely not be able to detect this as a block, but rather as a SSL error. This is due to the fact that a blocking page from the ISP would not match the SSL certificate of the requested page, what would then lead to the page loading error [95].

The content of a website which is called using the HTTPS protocol isn't transmitted as cleartext and should therefore not be able to get censored using keyword filtering blocking techniques. In case a website is blocked even if it is called using the encrypted HTTPS protocol this would probably indicate a block on a lower filtering level, e.g. through DNS censoring or blocking of hosting IP addresses. Most current internet browsers would call a website by default using HTTPS if possible, in case no specific protocol prefix is provided. Therefore, compared to the pages which are detected as blocked in the by the Blocked project, some amount might be accessible when called using HTTPS.

In the measurements from the OONI dataset 33% of the measurements used the HTTPS protocol.

As displayed in Table III, when comparing the rate of detected blocks and distinct blocked URLs, the rate of blocks is slightly smaller for the encrypted HTTPS protocol. This might indicate that using by using HTTPS some blocked websites might be accessible. However, the amount of tested domains using both protocols was too small in order to find perform an appropriate comparisons, which would allow to further analyse this aspect.

To further understand the effect of HTTPS on the blocking techniques applied by the ISPs, we compare the types of detected blocks protocol.

Figure 9 depicts the distribution of the detected blocking types per type of application layer. For websites accesses

Table III
COMPARISON OF BLOCKING PER HTTP AND HTTPS PROTOCOL FORM THE OONI MEASUREMENTS

	Measurements	Detected Blocks	% Detected Blocks	Tested URLs	Blocked URLs	% Blocked URLs
HTTP	67,129	4,508	6.72%	1,319	456	34.57%
HTTPS	33,214	856	2.58%	510	161	31.57%

through HTTPS almost no *http-diff* blockings occur. This might also be due to the reason described by the Blocked project. In case the user is redirected to some block page by the ISP the SSL certificate cannot match and therefore a SSL error occur. This error might then be represented as a *http-failure* blockings. Interestingly, the share of DNS blocks is also considerable smaller for HTTPS sites compared to HTTP sites. This might be due to the adaptation of end-to-end encrypted DNS servers [96].

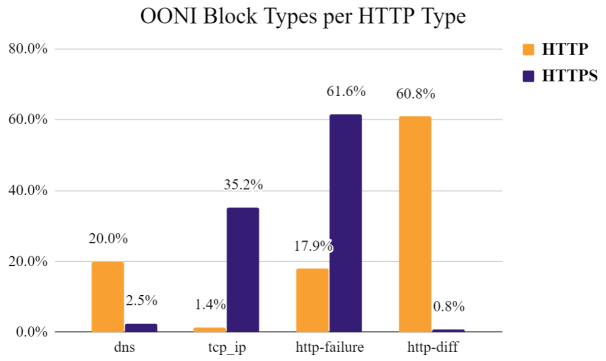

Figure 9. Distribution of blocking techniques in the detected blocks from the OONI Measurements per use of the encrypted HTTPS protocol and the cleartext HTTP protocol.

In Table XII and Table XIII we display the full comparison of the types of detected blocks per ISPs from the OONI measurements of encrypted and unencrypted websites.

C. High Court Order Blocks

The following table shows the implementation of the High Court order blocks as analysed by the ORG for every major ISP. Overall, 1106 URLs were tested. Of these, 46% were reachable, with a block being detected for 35% of them. The blocked column indicates an URL for which an UK High Court has determined a blocking to be necessary.

Nonetheless, as can be seen, many ISPs do not implement the High Court order blocks, with the column OK indicating an URL being reachable on a particular network. The blocked column indicates an URL being blocked on a network. ISPs such as Gigaclear, IFNL or LCHost seem to block no URLs at all, with a few errors being noted. Seeing that the number of errors, timeouts and DNS errors appear to be quite similar, it can be estimated that all of these ISPs encounter the same error, with none of them being responsible for it.

For AAISP no blocking rules are specified and therefore also no blocks can be detected. However, the network provider presents a special case among the ISPs, as they are the only major private ISP that guarantees blocking-free internet for their users and actively opposes internet censorship, taking into up potential legal ramifications for doing so. As this network is included in the Blocked projects measurements since 2014 and since then the block detection rules have been updated for many ISPs, it is likely that AAISP either does not display a block page for blocked websites, or indeed does not block websites at all.

The Sky network seems to be among the most strict, following nearly all ISP court order blocks. This is particularly distinct in comparison to the Sky-unfiltered network, which promises an unfiltered internet experience. However, the unfiltered Sky-network still allows 160 URLs to be reached.

Overall, it is clear, that High Court ordered blocks are not being implemented to the extent they should.

D. Categories of Blocked Websites

The displayed block pages from some of the ISPs provide information on the reason for a block. In the HTTP response bodies from the OONI measurement reports, it is also possible to view some of those blocking pages. One example blocking page, which does not rely on local HTML formatting resources is from the Web filtering as a service company *SafeDNS*. As depicted in Figure 10 on the blocking page the provider censored the page *peacefire.org*, as it was classified as a provider for proxy and anonymizing tools.

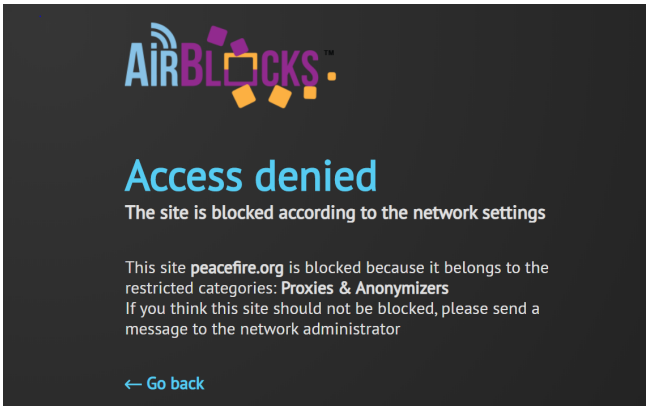

Figure 10. Example block page rebuild from the HTML file in the OONI measurement report: https://api.ooni.io/api/v1/measurement/temp-id-373456818

For some ISPs this blocking category is also extracted in the blocking rules in the Blocked project. For the Sky network this information is available for all blocked websites in our dataset. Therefore, it can be used to gain an initial idea about the content of the censored pages and whether this is in line with the censoring goals of UK policy makers. Figure 11 displays

Table IV
WEBSITE STATUS FOR COURT ORDERED BLOCKS PER ISP FROM THE BLOCKED ORG MEASUREMENTS.

ISP	ok	blocked	error	timeout	dnserror	sslerror	Tested Pages
AAISP	849	0	114	81	61	1	1106
BT	283	528	26	212	56	1	1106
BT-Strict	157	810	49	34	44	12	1106
BT-Unfiltered	199	676	24	178	29	0	1106
EE	727	61	118	90	102	8	1106
Gigaclear	826	0	113	70	75	1	1085
IFNL	730	0	109	75	51	1	966
Kingston Communications	827	0	112	68	75	1	1083
LCHost	828	0	113	67	75	1	1084
O2	646	61	97	201	100	1	1106
OpenDNS	642	130	72	53	58	1	956
Plusnet	307	555	10	161	72	1	1106
Plusnet-unfiltered	256	558	11	95	30	1	951
Sky	21	978	0	23	84	0	1106
Sky-unfiltered	161	832	16	16	35	16	1076
TalkTalk	351	660	15	76	0	4	1106
TalkTalk-unfiltered	265	712	45	74	9	1	1106
Three	662	145	91	104	103	1	1106
Uno	915	0	5	66	29	1	1016
VirginMedia	772	188	7	54	83	2	1106
VirginMedia-unfiltered	170	846	19	32	38	1	1106
Vodafone	67	904	2	30	97	6	1106
Zen Internet	822	0	113	74	75	1	1085
Percentages	46.33%	34.87%	5.17%	7.80%	5.57%	0.25%	100.00%

the categories for all censored websites in the Sky network. The large majority of those websites are blocked, as they are classified as pornography. This corresponds to the goals of protecting children from accessing adult material.

Services for increasing the anonymity on the internet are also being blocked. As described in Section V different methods exist for censoring websites. Some tools, such as proxy servers, can be used in order to circumvent censorship. Those tools achieve this by providing ways of using the internet in a more anonymous way. Therefore, by blocking such anonymizer services, it is also harder for internet users to keep their online privacy. Those services might be blocked with the goal of making it harder to gain access to content that violates copyrights, such as illegal movie streaming sites. Weapons, Dating, Drugs and Suicide sites most likely also correspond to the goal of restricting access for children to violent and adult content.

It is important to note that the sample of tested URLs in the measurements from the Blocked project are biased. As only the blocked websites are categorised, we are not able to distinguish the distribution of all tested websites. It might be that a by far larger share of websites with pornographic content has been submitted for testing and therefore the share of blocked sites in the dataset is larger than the remaining categories.

Therefore, the displayed distribution does not imply that pornographic content is blocked more strictly than websites from the remaining categories. Still, the provided names of blocked categories reveal the motivations behind the types of content that should be blocked.

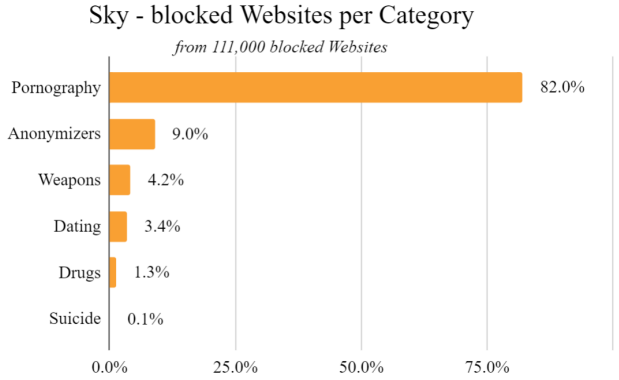

Figure 11. Distribution of categories for detected blocked websites in the Sky network from the Blocked project measurements.

As described in Section VII-C, it is also possible that websites with adult content are not blocked from accessing, but require an age verification from the internet user. Being able to distinguish between age verification pages and block pages would allow to answer the question which content is meant to be blocked for children and which for all citizens.

In the next section we use the collected textual content from those blocked websites to further investigate what type of content is actually displayed in the websites from the different categories.

IX. Analysis of the Scraped Data

For the analysis we focused on six categories that occurred most frequently. The categories are websites which are related to the following content: porn, drugs, dating, suicide, weapons or anonymizers.

The following subsections will put emphasis on the *description* of the categories and the data, followed by an explanation of the text cleaning process as well as data statistics. Afterwards, the most common words in each category and for each variable will be clustered and visualized via word clouds. Lastly, an analysis concerning the URLs is conducted and presented.

A. Data Description

The size of the categories varies a lot. Anonymizers are the biggest group with 10.490 observations, followed by porn with 10.020 websites and dating, which has almost 50% less entries (5.132). The smallest category is suicide with only 111 observations. Suicide is the only category that has too few entries in order to find patterns and significant findings but will still be briefly discussed. Each data frame has 12 variables. Some have a high number of missing values such as rankwords which are always close to 100% or have unserviceable content such as *X* containing only 0. But the datasets also contain variables with high informational content such as the variables *url*, *title*, *description* and *keywords*.

The variable *url* does not contain any missing values by nature because the *url* was given and scraped from the meta data on the websites. Additionally, it does not need any text cleaning. String manipulations that have been applied will be discussed in the analysis of the URLs in IX-E2.

As observed in Table V, the variables *title*, *description* and *keywords* contain many missing values. Especially, the *keywords* contain on average 72% missing values. Even though *title* has the least missing values, it still lacks content on 31% of the entries. However, it is less than most of the other variables. Nevertheless, the exclusion of the missing values from the dataset reduces the data on average by about 50%. But since the dataset size of the most categories is fairly large, the remaining data is still sufficient for data analysis.

Even though the variable *keywords* contains many missing values, its informational value is high. On average across categories, the variable *keywords* contains 21.059 words in total with 6.483 unique words. As shown in Table VI, the number of unique words is almost as high as the number of unique words of the variable *description*. Although, the variable *description* has a total word amount of 30.290 words. Thus, *description* has a higher number of repetitive words than *keywords* and *title*.

Table V
Number of observations and missing values of the most relevant variables for analysis per category

Category	No. Obs.	Missing Values		
		Description	***Title***	***Keywords***
Porn	10.020	5.345 (53%)	3.806 (38%)	7.324 (73%)
Drugs	1.020	527 (52%)	333 (33%)	762 (75%)
Dating	5.132	2.002 (39%)	1.338 (26%)	2.988 (58%)
Weapons	4.205	2.156 (51%)	947 (23%)	2.876 (68%)
Suicide	111	63 (57%)	30 (27%)	88 (79%)
Anonymizers	10.490	6.081 (58%)	4.229 (40%)	8.171 (78%)
Average	5.163	2.696 (52%)	1.781 (31%)	3.702 (72%)

B. Text Cleaning

For the text cleaning the R package stringr is used. The following manipulation and transformation techniques are applied:

- All letters are transformed to lowercase letters
- German umlauts (ä, ö, ü) as well as ß are manually transformed to ae, oe, ue and ss (especially relevant for the weapon dataset)
- Characters that are not from a-z are excluded
- White spaces are added behind all commas (relevant for keywords)
- English stop words as well as the word ”also” (manually) are removed
- Double white spaces are reduced to one white space
- Spaces in the beginning and the end of a sequence are removed
- Duplicate words within each entry are removed (relevant for the cluster analysis)
- All strings that are empty after text cleaning are removed

Table VII and VIII show the median number of characters (with white spaces) and the median number of words – both before and after cleaning.

The median is a more suitable measure of central tendency than the mean since the distribution of number of characters and words contains a lot of upper extreme values and is very/strongly right skewed. For instance, the keywords of the category weapons have a maximum length of 15.237 characters (with white space). However, the 3rd quartile is at 294. This results in a mean value of 292 but a median value of 149 since it is robust towards extreme values.

As observed in Table VII, the number of characters (with white spaces) is much lower for *title* than for *description* and *keywords*. However, the average median length reduction due to text cleaning is 13%. The more entries a category has, the higher the reduction of characters (with white spaces). For

Table VI
THE TOTAL NUMBER OF ALL AND UNIQUE WORDS PER VARIABLE AND PER CATEGORY (AFTER CLEANING)

Category	**Description**		**Title**		**Keywords**	
	Total	***Unique***	***Total***	***Unique***	***Total***	***Unique***
Porn	57.364	12.193 (-79%)	30.600	7.926 (-74%)	34.808	10.851 (-69%)
Drugs	6.795	3.048 (-55%)	4.145	2.049 (-51%)	5.228	2.838 (-46%)
Dating	43.117	8.574 (-80%)	20.374	5.322 (-74%)	29.080	6.929 (-76%)
Weapons	27.612	8.324 (-70%)	15.653	5.738 (-63%)	30.151	11.866 (-61%)
Suicide	700	525 (-25%)	371	246 (-34%)	353	293 (-17%)
Anonymizers	46.149	7.665 (-83%)	24.968	5.723 (-77%)	26.733	6.123 (-77%)
Average	30.290	6.722 (-65%)	16.019	4.501 (-62%)	21.059	6.483 (-58%)

Table VII
THE MEDIAN NUMBER OF CHARACTERS (WITH WHITE SPACES) IN A SEQUENCE BEFORE AND AFTER TEXT CLEANING PER VARIABLE AND CATEGORY

Categories	**Description**		**Title**		**Keywords**	
	Before	***After***	***Before***	***After***	***Before***	***After***
Porn	139	83 (-40%)	45	33 (-27%)	104	52 (-50%)
Drugs	144	103 (-28%)	47	36 (-23%)	148	107 (-28%)
Dating	150	98 (-35%)	50	39 (-22%)	139	71 (-49%)
Weapons	132	95 (-28%)	37	31 (-16%)	146	106 (-27%)
Suicide	142	102 (-28%)	27	24 (-11%)	109	105 (-4%)
Anonymizers	127	72 (-43%)	34	24 (-29%)	111	59 (-47%)
Average	139	92 (-34%)	40	31 (-21%)	126	83 (-34%)

Table VIII
THE MEDIAN NUMBER OF WORDS BEFORE AND AFTER TEXT CLEANING PER VARIABLE AND CATEGORY

Category	**Description**		**Title**		**Keywords**	
	Before	***After***	***Before***	***After***	***Before***	***After***
Porn	19	13 (-32%)	7	5 (-29%)	11	8 (-27%)
Drugs	20	14 (-30%)	7	5 (-29%)	17	14 (-18%)
Dating	22	14 (-36%)	8	6 (-25%)	15	10 (-33%)
Weapons	18	13 (-28%)	5	4 (-20%)	15	14 (-7%)
Suicide	18	13 (-28%)	4	4 (0%)	13	15 (+15%)
Anonymizers	18	11 (-39%)	4	4 (0%)	12	9 (-25%)
Average	19	13 (-32%)	6	5 (-17%)	14	12 (-16%)

instance, the categories porn, dating and anonymizers have the most observations and the highest reduction percentages.

Analogously, this can be observed for the median number of words before and after cleaning in Table VIII. Interestingly, a positive value can be observed for *keywords* for the category suicide, see Table VIII. When cleaning the texts, words like “pro-anorexia” become “pro” and ”anorexia”. Another example would be unknown characters that have not been specified or transformed unlike the German umlauts such as “Québec”. “Québec” becomes “Qu” and ”bec”. Hence, one word is split into two or possibly even more words. It is very complex to find all exceptions of characters that do not occur in the a-z. Moreover, after doing a random manual check, it does not have a strong impact on the results. Suicide is the only dataset influenced by it as a result of its size. However, as mentioned before the dataset regarding blocks due to suicidal content will only be analyzed manually and hence, does not pose a problem.

As stated in the overview of the conducted data manipulation, the transformation of the German umlauts, especially of ü, was very important for the category weapons which contains a lot of words like “schützenverein” for instance. Not transforming them has a strong impact on the word count (“schützenverein” becomes “sch” and “tzenverein”) since it occurs quite often, but also on the clustering analysis and visualization because the word cannot be fully presented and mixed up with similar words like “schützengesellschaft” or

rather “sch” and “tzengesellschaft”.

The reason for the addition of whitespaces after commas is that many words are listed like “marry,avoid,snogme,snogfm,snogs,marries,avoids,social network,social,meet people (...)”[5]. When counting the number of words, this is counted as one word. Thus, without the transformation, the number of words, especially for keywords, increased almost every time after cleaning.

C. Cluster Analysis

1) Distance Matrix: Generally distance matrices only accept numeric input variables. However, in our case we have textual data. The stringdist R package offers functions that compute distance matrices based on similarities of strings.[6]

In order to measure the similarity of the given textual sequences, an edit-based distance measure approach is chosen. The generalized or rather original Damerau-Levenshtein (dl) distance is an extension of the Levenshtein distance [97]. As well as the Levenshtein distance, the Damerau-Levenshtein distance measures the minimum amount of character edits that are needed to transform one word into another. The difference among those two distance measures are the possible edit operations. Not only insertions, deletions and substitutions of single characters are feasible but also the transposition of two adjacent characters. [98]

$$d_{dl}(s,t) = \begin{cases} 0, \text{if } s = t = \epsilon. \\ min\{ \\ d_{dl}(s, t_{1:|t|-1}) + w_1, \\ d_{dl}(s_{1:|s|-1}, t) + w_2, \\ d_{dl}(s_{1:|s|-1}, t_{1:|t|-1}) + [1 - \delta(s_{|s|}, t_{|t|})]w_3, \\ min_{(i,j)\in\wedge} d_{dl}(s_{1:i-1}, t_{1:j-1}) + [(|s| - i) \\ +(|t| - j) - 1]w_4\}, \text{otherwise.} \end{cases} \tag{1}$$

with a minimization (in the last line) over

$$\wedge = \{(i,j) \in \{1,\dots,|s|\} \times \{1,\dots,|t|\} : s_s = t_j, s_i = t_{|t|}\}.$$

with $w_1 = w_2 = w_3 = w_4 = 1$. w_1, w_2 and w_3 are the non-negative penalties for the edit operations insertion, deletion and substitution and w_4 is the penalty for transposition when transforming t to s. ϵ presents an empty string and d the distance function [97].

The maximal distance between two string s and t is $max\{|s|, |t|\}$ [97]. As shown in equation 1 the minimal distance is 0 if the strings s and t are exactly the same.

Another approach for calculating distances for textual data are heuristic distance measures like the Jaro–Winkler distance. However, the Jaro-Winkler distance is most efficient for short sequences like names for instance. As observed in Table VIII, the median number of words for *description* is 92 and thus, does not qualify as a short string.

2) Cluster Algorithm: Since our analysis is a new approach of grouping blocked websites and has not been performed before, we do not have any indication about the number of the clusters per category. Hence, a hierarchical clustering approach was chosen over a clustering approach where the number of clusters has to be determined before the analysis as necessary for the c-means algorithm for instance. The results of the group-building algorithm can be visualized with a dendrogram which presents the distance between clusters on the y-axis and indices of the observations on the x-axis. The dendrogram gives an indication of how many clusters to choose. [99] The results can be ambiguous. In general, the trees should be cut when the distance between clusters is maximized while the distance within clusters is minimized [100]. However, the interpretation primarily has to follow the explanatory purpose.

As a grouping algorithm, Ward (ward.D) was implemented which is one of the most used grouping algorithm. Ward merges observations based on the smallest increase of the total within-cluster variance. [99]

The results of distance based grouping algorithm such as complete linkage (compares largest distances) and single linkage (merges closest objects) were quite poor because the distances are very similar [99]. The median distance of the porn category for instance is 98. This means in average 98 edit operations have to be performed in order to change string s to string t.

3) Application: For the application of the cluster analysis, the variable *description* was used. Besides having the most words in all categories (except weapons[7]), *description* has the lowest percentages of unique words compared to the number of total words, see Table VI. It is favourable if the number of repeated words in comparison to the number of total words is high in order to find and group similar entries. Unfortunately, synonyms cannot be detected.

After checking with the dendrogram and the average silhouette width per cluster statistics, for all categories the selection of three clusters was reasonable. Unfortunately, many clusters have negative silhouette width which indicates that they might be assigned to the wrong cluster. For all clusters word clouds are created. However, only the word clouds with positive silhouette width can be discussed and interpreted.

The overall silhouette widths per category tend to increase with the total number of observations per category, see Table V and Table IX. In contrast, cluster with less observations perform better in general. However, the visualisation in form of a word cloud is less effective. For instance, Figure 12 (bottom) represents cluster 3 from a porn category has an average silhouette width of 0.8 which is very high, but the cluster has too few observations to form a proper word cloud. But still, they contribute interesting findings and thus, add value to the analysis.

[5]exemplary excerpt of an entry of the variable *keywords* from the dataset dating

[6]https://cran.r-project.org/web/packages/stringdist/index.html

[7]*description* has 27.612 entries; the variable *keywords* has slightly more with 30.151 observations

Table IX
CHOSEN NUMBER OF CLUSTERS WITH THE ASSIGNED OBSERVATIONS (OBS.) AND THE AVERAGE SILHOUETTE WIDTH PER CLUSTER (AVG. SILWIDTHS)

No. of	Porn		Drugs		Dating		Weapons		Anonymizers	
Clusters	*Ob.*	*Avg. Silwidths*	*Obs.*	*Avg. Silwidths*	*Obs.*	*Avg. Silwidths*	*Obs.*	*Avg. Silwidths*	*Obs.*	*Avg. Silwidths*
1	1973	-0.17	404	0.63	1203	-0.01	404	0.63	932	0.52
2	1486	0.47	1012	-0.09	915	-0.14	1012	-0.09	1191	-0.15
3	232	0.80	101	-0.15	266	0.72	101	-0.15	348	0.80

D. Word Clouds

In order to be shown in the word cloud, a word has to appear at least twice. If it occurs less, it is not visualized.

The word cloud uses the cleaned *description* of each cluster per category as an input and transforms it to a term document matrix which is in turn transformed to a matrix. With help of the R package wordcloud[8], the words inclusive their frequencies can be derived from the matrix and can be visualized.

E. Interpretation

The word "information" occurs in the categories porn, drugs, suicide and anonymizers in the top 20 of the most frequent words of the variable *description* (see Table X). This might indicate a website for help or support, especially for thoughts of suicide or drug abuse. However, it could also be information on how to commit suicide, where to buy drugs etc. The following section digs deeper into the most striking findings of the category porn, weapons, drugs and suicide and investigates what kind of websites are blocked. We use the URLs, tables but also the formed word clouds that characterize the difference between blocked websites.

1) Description Variable: As presented in Table IX the average silhouette widths for cluster 2 und 3 in the category porn are 0.47 and 0.8. Especially, the silhouette width of cluster 3 indicates that the observations are well clustered which means that the content of the description is very similar. Even though, the cluster is (with 232 observations) much smaller than the other two clusters, it contains sufficient observations to further investigate and to give a valuable interpretation. The words "information" and "hope" for instance stand out as some of the most frequent words in the cluster. Both are not typical words for porn websites and rather indicate other types of websites. Opening three randomly selected websites contained in cluster 3 supports the thesis.

113 of the 232 entries feature a nearly identical description: "website sale [name of website] first best source information re looking general topics expect find hope searching" – this finding also explains analogously the composition of websites of cluster 3 for the anonymizers category. Those entries are domains that are offered for sale. However, the names do not necessarily indicate any pornographic activities: http://www.officialbookmark.info, http://www.piercing-bewertung.de and http://www.vintagelampstore.info. By randomly checking some websites that actually indicate pornographic content,

[8]https://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf

Figure 12. Porn cluster 2 (top) and porn cluster 3 (bottom).

we observed unavailable websites or a 404 response for the most part. An example of a website that offers a domain for sale is given in Figure 13.

Figure 13. Piercing-bewertung.de is an exemplary website that is for sale of cluster 2 in the porn category.

Table X
30 MOST FREQUENT WORDS FOR THE VARIABLE DESCRIPTION PER CATEGORY

Rank	Porn		Drugs		Dating		Weapons		Suicide		Anonymizers	
	Word	*Freq*	*Word*	*Freq*	*Word*	*Freq*	*Word*	*Freq*	*Word*	*Freq*	*Word*	*Freq*
1	porn	973	cannabis	96	dating	1140	gun	186	suicide	11	free	924
2	free	871	online	67	site	725	und	166	information	7	looking	820
3	videos	833	marijuana	64	free	691	online	137	life	7	online	673
4	sex	813	information	50	singles	627	firearms	137	assisted	7	best	590
5	com	664	drug	41	online	613	accessories	136	site	6	find	474
6	best	612	best	39	find	503	shooting	121	tips	5	download	464
7	looking	435	buy	38	meet	414	shop	112	die	4	proxy	451
8	find	396	cbd	38	love	360	guns	107	end	4	source	447
9	girls	388	products	35	com	307	best	106	exit	4	links	442
10	movies	356	shop	34	single	296	quality	102	euthanasia	4	information	435
11	site	346	free	31	chat	295	com	99	eating	3	related	425
12	xxx	346	medical	30	women	280	custom	97	may	3	first	417
13	tube	294	hemp	30	join	272	hunting	96	leben	3	general	417
14	hot	284	quality	27	looking	241	knives	94	sterben	3	see	415
15	watch	279	including	25	men	217	der	93	self	3	topics	412
16	first	273	und	25	date	207	world	90	membership	3	searching	403
17	adult	266	high	24	today	206	products	90	anorexie	3	hope	402
18	source	264	headshop	23	people	205	rifles	83	boulimie	3	movies	401
19	information	262	com	23	best	204	sale	82	les	3	watch	401
20	online	259	find	23	now	185	fuer	82	pour	3	expect	398
21	sexy	259	can	22	service	177	one	70	sur	3	com	383
22	porno	244	grow	21	und	152	buy	69	commit	3	sale	351
23	hope	240	seeds	21	world	145	rifle	68	don	3	website	347
24	expect	239	get	19	marriage	141	des	67	drop	3	domain	263
25	general	239	oil	19	website	138	site	66	drugs	3	web	258
26	searching	239	world	18	new	135	schuetzenverein	65	everything	3	unblock	251
27	topics	239	sale	18	sites	124	find	65	hanging	3	may	218
28	sale	232	links	18	register	123	den	64	hope	3	websites	212
29	video	215	legal	18	friends	120	free	62	lost	3	can	194
30	see	212	order	17	community	118	new	61	reading	3	music	191

In contrast, cluster 3 seems to contain stereotypical porn websites. The most frequent words are, among others, “free”, “porn”, “videos”, “sex”. A block might be argued for on the basis of child protection. However, by looking closer at the *description* of the porn websites of cluster 2, many Japanese and Chinese websites occur. They are mixed starting with betting websites, websites offering part or full-time jobs in the sex industry, fan websites with images of Shotacon/Kemoshota[9] but also websites of a plastic surgeon for instance (see exemplary screenshots in Figure 14). We checked the Japanese as well as the Chinese websites that have been blocked without obvious pornographic content at first sight with native speakers.

[9]Shotacon is a Japanese slang word for the attraction to young boys. The images are visualized in manga style [101]; Kemoshota is similar but in form of animals as the object of attraction.

Based on native speaker expertise, some contain content that fall under the goal of restricting access to adult content, such as the website http://www.s-gcn.com which is a kind of club where man can find pleasure through watching women live with the option of paying them for a joined service. The website http://www.yanchana.net is also blocked. The presumed reason is that the UK blocks websites that are associated with the keywords Shotacon or similar. At the time of our analysis the site only depicted regular manga images (see Figure 15). However, websites associated with this keyword might depict other types of content which would lead to blocking of all pages associated with Shotacon.

Another website we examined is the website of a plastic surgeon that is advertised for being especially talented in surgical gender reassignments. However, he is also offering to

Figure 14. Exemplary websites from the porn cluster 2 - a Japanese website of plastic surgeon (top) and a Japanese website for Shotacon/Kemoshota (bottom).

perform other plastic surgeries. We could not determine any obious blocking reasons at first sight. It did not contain any apparent harmful, disturbing or pornographic content during our investigation.

Figure 15. Exemplary screenshots of the website http://www.yanchana.net.

The exemplary websites discussed above are just a few of many examples that exemplify overblocking in the UK.

The weapons category also provides a plethora of questionable blocks. In total, the cleaned dataset of the category weapons (with 1517 entries) contains 153 websites of shooting clubs ("Schützenvereine"), 44 archery websites, 47 hunting websites and 84 websites that are related to shooting sports. The websites are – with very few exceptions from Austria and Switzerland – from Germany. The words "Schützengesellschaft", "Schießverbund", "Bogensport", "Jagd" or similar occur in their website description. Especially, the "Schützenverein" websites that we randomly checked, did not show any evidence of harmful, aggressive or violent content. The tradition seems to be the focus of these websites (see Figure 16 (top)).

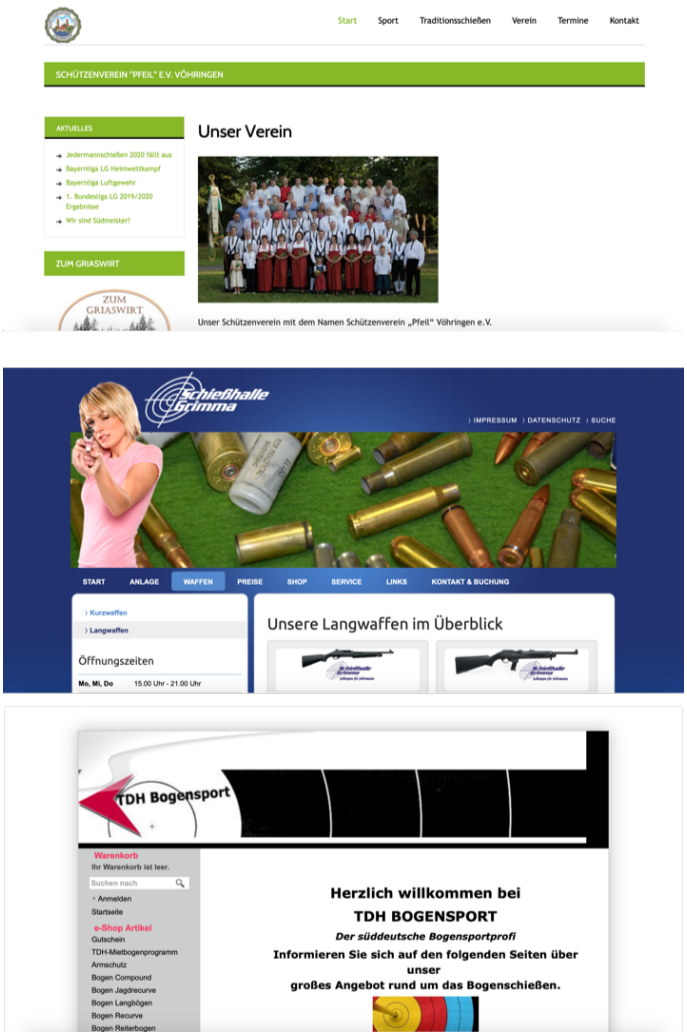

Figure 16. Exemplary screenshots of a Schützenverein (top), a shooting sport (middle) and an archery club (bottom) website.

Weapons occur on some photos but are usually not in the focus of attention. In contrast, the shooting sports and hunting websites present weapons that are also offered for sales and advertised very prominently. We assume that keywords such as "Munition", "Pistole" or also specific weapons name such as "Heckler & Koch SFP9 SF", "Rep. Büchse Mercury Urban Sniper" might cause the block or filter. The archery websites are websites of sport clubs or of stores that

sell bows and equipment. While bows might be classified as weapons, they like many other sports such as fencing, aikido or javelin, among others, are part of a known and established sport. The blocking of sports related websites just for the sake of weapons being involved would interfere strongly with the personal freedom of individuals. Endangerment of youth should not be suspected from an Olympic sport. Furthermore, by randomly checking other websites, we found websites that offer deep sea fishing, websites that offer personal protection such as face masks, pepper sprays etc., a blog about transgenderism/feminism and the website of the computer game Battlefield. For many of the examined websites it is to be doubted that young people could be influenced negatively or laws be broken, thus personal freedoms are being restricted without further justification. Conclusively, the category weapons contains also many overblocks.

The exact blocking rules and implementations are unknown. The British Board for Film Classification (BBFC) gives an overview of rules for blocking. It is a non-profit and independent regulator that provides standards on how the age can be verified before accessing a website [55]. These regulations are based on the Digital Economy Act 2017 (DEA) described in III-E6. The BBFC creates and maintains an access control tool called the BBFC Mobile Classification Framework. The framework sets out guidelines for commercial content blocking and minimum age limits for content on mobile networks. It is intended to stop minors from accessing a website via mobile phones. The guidelines that the Classification Framework is based on were derived from UK law as well as public consultations and credible media effects researches. However, the rules whether a website should contain an access control do not seem to be transparent, weirdly specific and randomly picked. For instance, websites could be subject to age and access limitations if they contain a "repeated / aggressive use of 'cunt'". [56] The word "cunt" is the only listed word here, no other words are listed. Nevertheless, the Classification Framework can give an indication for why some websites are blocked since the guideline is based on law and constantly adapted and improved.

For websites referring to drugs and suicide the guideline says that the promotion, instructions, portrayal and glamorisation should lead to access control [56].

In case of the websites of drugs this results in rehabilitation center and educational websites having an access control such as http://www.psychedelic-education.org because they tend to portray the drugs but also the consequences of misusing them.

Some websites that were blocked or access controlled due to suicidal content contain fitness programs such as https://www.activefit.org/ (see Figure 17).

Moreover, the websites that offer support and share stories about their self-injury processes are either block or access controlled because – as well as the websites offering help for drugs additions – the websites portray illnesses like self-injury that sometimes co-occur with suicidal thoughts (e.g. http://www.psyke.org).

About 30% of the access controlled websites have the URL ending .org. The URL endings will be discussed in detail in the section IX-E2.

However, besides questionable blocks, also websites were blocked/access controlled with information on how to commit suicide which is the originally goal of the guideline that is provided by BBFC.

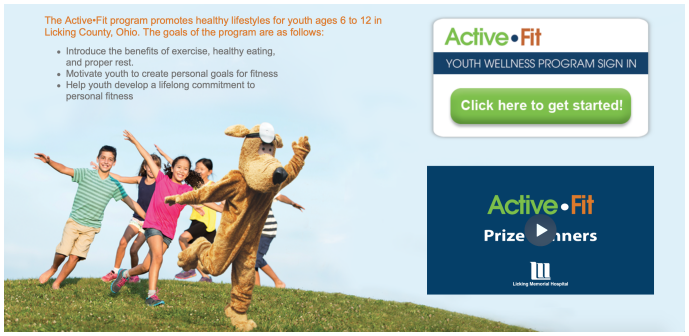

Figure 17. Exemplary screenshot of a fitness website for children between 6 and 12 with sample exercise for instance.

Overall, we showed across several categories many examples that indicated – from the current perspective (content might always change) – overblocking. The consequences and the results of the interpretation will be discussed in detail in section X.

2) URLs: The URLs contained in the dataset also give valuable insights into the content and origin by looking at the top-level domain (TLD), see Table IX-E2 (on the top of the next page). Top-level domains enable one to put a domain into one of three categories: countries, multi-organizations and categories [10]. This allows for country-specific content localization, and thus gives the information, content originating from which countries is being blocked.

The most common top-level domain is by far *.com*, independent of the blocking category associated with it. Such generic top-level domains are assigned free of origin country and designate a commercial entity, showing that the content accessible through this website is most likely hosted by a commercial entity. A *.com* TLD does not contain information on the country the content is hosted in. Another commonly blocked TLD is the *.de* domain, which indicates a German TLD. It is the second most frequently found top-level domain in the dating and weapons category, further undermining the observation that many of the websites blocked under the weapons moniker are German shooting clubs.

Of particular interest is the *.org* TLD found with high frequency in the suicide categories. This TLD is commonly used and was intended for non-profit entities, with it being opened for public registration only in August 2019 [102]. A plethora of NGOs make use of the .org TLD, among them many that have made it their goal to prevent drug abuse and help addicts.

[10]https://tools.ietf.org/html/rfc920

Table XI
10 MOST FREQUENT URLS PER CATEGORY

Rank	Porn		Drugs		Dating		Weapons		Suicide		Anonymizers	
	Word	***Freq***	***Word***	***Freq***	***Word***	***Freq***	***Word***	***Freq***	***Word***	***Freq***	***Word***	***Freq***
1	com	6116	com	474	com	2996	com	2107	com	39	com	3866
2	net	825	org	153	de	305	de	639	org	22	net	1188
3	ru	485	uk	68	net	242	org	225	net	9	org	671
4	org	203	de	54	uk	222	ch	154	ch	5	ru	339
5	info	191	net	41	org	136	net	144	au	4	tv	318
6	tk	133	nl	21	ru	105	uk	138	html	4	me	266
7	xxx	98	ru	20	fr	81	html	119	uk	4	to	254
8	de	95	shtml	20	jp	76	htm	98	be	3	eu	230
9	ga	94	ca	16	dating	61	it	78	de	3	info	198
10	uk	94	html	16	nl	49	dk	58	it	3	co	181

X. DISCUSSION AND LIMITATIONS

The analysis performed clearly showed the extent of internet censorship in the United Kingdom. It is evident that blocking is widespread and far reaching, with overblocking and false positives being a regular occurrence. It sheds significant light on the inefficiencies of blocking in the way it is implemented in the UK, questioning the necessity and legal bases for it at the same time. However, it is limited in its scope and the data collected and does not claim to be comprehensive in its applicability.

Of the ISPs looked at in this paper, most are concerned with providing internet to private citizens. Other ISPs are mostly concerned with commercial traffic for large companies and institutions exist and for the most part do not implement traffic filtering. It seems unlikely that filtering will be implemented for these entities, violating net neutrality standards according to which all traffic should be considered as equal. Besides, filtering can be very different depending on where you live as a result of different ISPs offering service in different regions. AAISP for example is one of the few British ISPs only offering unfiltered internet access, but are limited to certain regions, while council estates sometimes only have access to BT provided internet [103].

Notable here is the fact that blocking is often the default setting on broadband packages, with unfiltered connections also being available from ISPs such as BT and Sky upon specific requests. The option to disable filtering is often not advertised and it is to be doubted whether a significant percentage of people with censored internet are aware of their internet being filtered. Of the URLs used for the scraping analysis, all of them were found on the Sky network, limiting the potential customers affected by blocking to at most 6.1 million websites in the UK [104].

Furthermore, blocking can vary drastically between ISPs, with the same hypothetical user having a different set of blocked websites based on their choice of ISP. It is not possible to determine the exact blocking methods used by each ISP, which only compounds one of the main problems of ISP based web blocking in the UK, the legal overblocking. With ISP unable to be prosecuted for blocking access to legitimate and harmless content, no clear incentive for overblock reduction exists. With multiple businesses already having been blocked as a result of overblocking by an ISP, the ramifications for web business are apparent, as they depend on web traffic for business purpose.

Even with web filtering in place, it is possible for an user to circumvent censorship and access content supposed to be unreachable. Making use of the encrypted HTTPS protocol, which is already implemented by most websites, allows one to circumvent some of the used blocking mechanisms. VPNs and Proxies can also aid users in circumventing censorship, although efforts are being made to block access to these services too. It can be assumed that a majority of citizens would be knowledgeable enough to find a way around their ISP censorship, but are either not aware of it or do not make the effort to do so, leaving them at risk for further restriction of their right to privacy.

All ISPs base their blocking implementation on blocking lists that they either create themselves or are given by secondary institutions such as the IWF. This leads to a lack of overview of who is in control of what is being blocked. Furthermore, it removes accountability from a central institution and ensures that no one entity is solely responsible for a website block appearing. Additionally, with the High Court order blocks, blocks can be legally enforced by rightsholders, but are subject to a costly legal process. This ensures that monetary resources can directly affect web censorship and excludes penniless rightsholders from enforcing their rights to personal ownership on the internet. It might be possible in the future for more stakeholders to enforce legal blocks thanks to new legal processes, which might allow faster and lower cost process of enforcing legal blocks.

All analyses performed throughout our research are subject to an introduced bias as a result of chosen methods and tools. The scraped websites for example were cleaned in such a way that only letters of the alphabet (a-z) were left,

leaving out, among others, Russian, Chinese, Japanese and Arabic information in many instances. This was done for reasons of standardization and language understanding. Thus, websites from countries not using the Roman alphabet are being blocked too but were not part of further analyses in this research. However, some contained the letter combination "ae" and therefore, were kept. Natural language processing tools might help in incorporating foreign websites with different characters too through automated translation, but were infeasible in our case as a result of scope limitations. The distance matrix used above also does not recognize synonyms and words of the same meaning in different languages, nor does it evaluate contextual differences. However, it still gives a good indication of a website's main content and shows questionable blocks such as that of the German "Schützenvereine".

Further research should put their focus on improving clustering in regard to the limitations named above. With enough computing power, web scraping might also be done with the help of a neural network, such as the Dragnet implementation based on a paper by Matthew Peters and Dan Lecocq [105]. This would result in a larger dataset too, improving statistical validity.

XI. Outlook and Conclusion

The topic of internet censorship might be associated with countries ruled by oppressive regimes. As shown in this work, democratic regimes such as that in the UK are also keen on controlling their citizen's internet, restricting personal freedom under the guise of increased security and harm reduction.

Our research provides an oversight of the foundations for internet censoring in the UK and depicts the extent of censorship. Still, further research is needed. Using our methodologies it is not known in detail, which ISP makes use of which technical implementation, just as well as it is unknown how many URLs are being blocked at any point by any ISP.

Given the goals of limiting child access to adult content and restricting access to copyrighted materials, services for increasing the anonymity on the internet are blocked. It seems obvious that circumvention of blocking is to be prevented too, ensuring the ISP and, therefore, the British government to have control over the content UK citizens can access over the internet.

Although the measures for censoring the internet are widespread, their effectiveness is still limited. As the discussion above shows, the circumvention of internet censorship in the UK is possible, but requires effort and knowledge from the end user. With overblocking being legal and occurring frequently, an overview of blocked websites not being available and an existing infrastructure for complete web censorship in place, it would be of ease for a ruling British government to exploit internet censorship for their own benefit. Additionally, while it is easy to attribute a lot to government interests, in particular that of former Premier Ministers Cameron and Theresa May, public support was always behind the measures against child pornography, with further freedoms being taken away in the name of child protection and online safety. It remains to be seen whether a balance between freedom and security can be struck.

Looking ahead, the effects of the Brexit might also be felt in the UK internet sphere. At the moment, European data and privacy rights also apply to UK citizens, protecting them from widespread governmental data collection and retention and extensive public tracking. With the UK leaving the European Union, these rights might no longer apply, leaving UK citizens without legal protection from a government intent on keeping track of every citizen's internet activity. Further research on this development is needed in order to keep shedding light on the state of internet censorship in the UK.

REFERENCES

[1] ITU. (2019) Measuring digital development: Facts and figures 2019. [Online]. Available: https://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx

[2] J. Manyika and C. Roxburgh, "The great transformer: The impact of the internet on economic growth and prosperity," *McKinsey Global Institute*, vol. 1, 2011.

[3] B. Liang and H. Lu, "Internet development, censorship, and cyber crimes in china," *Journal of Contemporary Criminal Justice*, vol. 26, no. 1, pp. 103–120, 2010.

[4] R. Rosenzweig, "Censorship in research and scholarship," 2001.

[5] J. M. Coetzee and J. M. Coetzee, *Giving offense: Essays on censorship*. University of Chicago Press, 1996.

[6] First Amendment, *US Constitution*, 1791.

[7] K. Stamm, "Das bundesverfassungsgericht und die meinungsfreiheit," *Aus Politik und Zeitgeschichte, Bd*, vol. 37, no. 38, pp. 16–25, 2001.

[8] K. D. Ewing, "The human rights act and parliamentary democracy," 1999.

[9] M. Heins, *Sex, sin, and blasphemy: A guide to America's censorship wars*. New Press New York, 1993.

[10] M. Newth, "The long history of censorship. national library of norway, 2010. accessed on march 10, 2020," 2010. [Online]. Available: http://www.beaconforfreedom.org/liste.html?tid=415&art_id=475

[11] M. Mazur, "History of censorship timeline. march 21, 2018. accessed on november 20, 2019," 2019. [Online]. Available: https://www.preceden.com/timelines/174747-history-of-censorship-timeline

[12] LIS BD Network, "Types and definition of censorship in libraries. november 18, 2013. accessed on march 7, 2020," 2013. [Online]. Available: http://www.lisbdnet.com/types-of-censorship-in-libraries/

[13] C. Anthonissen, "18. the sounds of silence in the media: Censorship and self-censorship," *Handbook of communication in the public sphere*, vol. 4, p. 401, 2008.

[14] C. Woll, "Censorship under a military government. october 1, 2019. accessed on march 11, 2020," 2019. [Online]. Available: https://www.britannica.com/topic/censorship/Character-and-freedom

[15] B. J. Mauer, "Censorship is not all bad. september 3, 2016 (update: March 10, 2017). accessed on march 3, 2020," HuffPost, 2017. [Online]. Available: https://www.huffpost.com/entry/censorship-is-not-all-bad_b_9417646

[16] B. Miller, "11 chief pros and cons of internet censorship. september 4, 2015. accessed on march 3, 2020," GreenGarage, 2015. [Online]. Available: https://greengarageblog.org/11-chief-pros-and-cons-of-internet-censorship

[17] N. Regoli, "19 biggest pros and cons of censorship. july 23, 2019 (update: January , 28 2020). accessed on march 6, 2020," 2020. [Online]. Available: https://futureofworking.com/11-biggest-pros-and-cons-of-censorship/

[18] T. McIntyre, "Internet censorship in the united kingdom: National schemes and european norms," *Law, Policy and the Internet (Hart Publishing, 2018 Forthcoming)*, 2018.

[19] (2020) Internet access — meaning in the cambridge english dictionary. [Online]. Available: https://dictionary.cambridge.org/dictionary/english/internet-access

[20] (2016) What is internet access? - definition from techopedia. [Online]. Available: https://www.techopedia.com/definition/7776/internet-access

[21] O. for National Statistics. (2019) Internet access – households and individuals, great britain: 2019. [Online]. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/bulletins/internetaccesshouseholdsandindividuals/2019

[22] OpenRightsGroup. (2019) About blocked. [Online]. Available: https://www.blocked.org.uk/about

[23] NCAC. (2019) Internet filters. [Online]. Available: https://ncac.org/resource/internet-filters-2

[24] S. N. Hamade, "Internet filtering and censorship," in *Fifth International Conference on Information Technology: New Generations (itng 2008)*, April 2008, pp. 1081–1086.

[25] I. Society. (2017) Internet society perspectives on internet content blocking: An overview. [Online]. Available: https://www.internetsociety.org/resources/doc/2017/internet-content-blocking/

[26] M. Burgess. (2016) More than 1,000 uk schools found to be monitoring children with surveillance software. [Online]. Available: https://www.wired.co.uk/article/uk-school-surveillance-technology

[27] S. I. Centre. (2017) A guide to monitoring. [Online]. Available: https://www.saferinternet.org.uk/blog/guide-monitoring

[28] R. H'Obbes' Zakon, "Hobbes' internet timeline v7. 0," *Zacon Group*, pp. 1–32, 2004.

[29] CNN. (2001) How police smashed child porn club. [Online]. Available: http://edition.cnn.com/2001/WORLD/europe/UK/02/13/paedophile.police/index.html

[30] K. Jaishankar, D. Halder, and S. Ramdoss, "Pedophilia, pornography, and stalking: Analyzing child victimization on the internet," *Crimes of the Internet*, pp. 28–42, 2009.

[31] U. Government. (2003) Sexual offences act 2003. [Online]. Available: http://www.legislation.gov.uk/ukpga/2003/42/contents

[32] I. M. C. Body. (2005) Imcb guide and classification framework for uk mobile operator commercial content services. [Online]. Available: https://ee.co.uk/content/dam/everything-everywhere/documents/IMCB%20Classification%20framework.pdf

[33] BBC. (2012) About sharing. [Online]. Available: https://www.bbc.com/news/technology-18071119

[34] A. News. (2007) Abc news live. [Online]. Available: https://abcnews.go.com/Blotter/story?id=3342018&page=1

[35] K. Sengupta. (2007) Police link suspects held over failed attacks - independent online edition ¿ crime. [Online]. Available: https://web.archive.org/web/20071001002250/http://news.independent.co.uk/uk/crime/article2737136.ece

[36] ORG. (2019) Counter-terrorism internet referral unit. [Online]. Available: https://wiki.openrightsgroup.org/wiki/Counter-Terrorism_Internet_Referral_Unit

[37] M. Jackson. (2018) Org warns many court ordered uk isp website blocks are in error - ispreview uk. [Online]. Available: https://www.ispreview.co.uk/index.php/2018/06/org-warns-many-court-ordered-uk-isp-website-blocks-are-in-error.html

[38] U. government. (2020) High court order blocks. [Online]. Available: http://www.ukispcourtorders.co.uk/

[39] E. Rosati. (2017) The ipkat. [Online]. Available: http://ipkitten.blogspot.com/2017/03/first-live-blocking-order-granted-in-uk.html

[40] Gov.uk. (2012) Parents asked if adult websites should be blocked. [Online]. Available: https://www.gov.uk/government/news/parents-asked-if-adult-websites-should-be-blocked

[41] E. Woollacott. (2013) Is the uk sleepwalking towards internet censorship? [Online]. Available: https://www.forbes.com/sites/emmawoollacott/2013/11/27/is-the-uk-sleepwalking-towards-internet-censorship/

[42] BBC. (2013) Pornography online: Lib dems. [Online]. Available: https://www.bbc.com/news/uk-24104110

[43] itv. (2013) Child protection web filters 'kept on' under new rules. [Online]. Available: https://www.itv.com/news/update/2013-11-16/child-protection-web-filters-to-be-kept-on-under-new-rules/

[44] J. Vincent. (2013) Abuse support and sex education sites blocked by isp's 'porn filters'. [Online]. Available: https://www.independent.co.uk/life-style/gadgets-and-tech/abuse-support-and-sex-education-sites-blocked-by-isps-porn-filters-9015389.html

[45] M. D. Smith. (2013) Some of the porn filters being offered by internet companies censor the wrong sites, as jim reed reports. [Online]. Available: https://www.bbc.com/news/uk-25430582

[46] reporters without Borders. (2014) Enemies of the internet. [Online]. Available: https://web.archive.org/web/20140312120731/http://12mars.rsf.org/2014-en/#slide2

[47] U. Government. (2016) Investigatory powers act 2016. [Online]. Available: http://www.legislation.gov.uk/ukpga/2016/25/contents

[48] U. government. (2016) Investigatory powers bill. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/473748/Factsheet-CD_Request_Filter.pdf

[49] ——. (2016) Investigatory powers bill - fact sheet. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/473745/Factsheet-Internet_Connection_Records.pdf

[50] A. Griffin. (2016) Everyone who can now see your entire internet history, including the taxman, dwp and food standards agency. [Online]. Available: https://www.independent.co.uk/life-style/gadgets-and-tech/news/investigatory-powers-bill-act-snoopers-charter-browsing-history-what-does-it-mean-7436251.html

[51] O. Bowcott. (2016) Eu's highest court delivers blow to uk snooper's charter. [Online]. Available: https://www.theguardian.com/law/2016/dec/21/eus-highest-court-delivers-blow-to-uk-snoopers-charter

[52] U. government. (2017) Digital economy act 2017. [Online]. Available: https://www.legislation.gov.uk/ukpga/2017/30/pdfs/ukpga_20170030_en.pdf

[53] J. Waterson. (2019) Uk drops plans for online pornography age verification system. [Online]. Available: https://www.theguardian.com/culture/2019/oct/16/uk-drops-plans-for-online-pornography-age-verification-system

[54] Y. Akdeniz, "Internet content regulation: Uk government and the control of internet content," *Computer Law & Security Review*, vol. 17, no. 5, pp. 303–317, 2001.

[55] B. B. for Film Classification. (2013) Mobile content. [Online]. Available: https://bbfc.co.uk/what-classification/mobile-content

[56] BBFC. (2019) Framework. [Online]. Available: https://bbfc.co.uk/what-classification/mobile-content/framework

[57] A. Travis. (2000) Watchdog moves to curb racist websites. [Online]. Available: https://www.theguardian.com/technology/2000/jan/26/internet.raceintheuk

[58] E. B. Laidlaw, "The responsibilities of free speech regulators: An analysis of the internet watch foundation," *International Journal of Law and Information Technology*, vol. 20, no. 4, pp. 312–345, 2012.

[59] I. W. Foundation. (2018) What we do. [Online]. Available: https://www.iwf.org.uk/report/2018-annual-report

[60] D. R. Clayton. (2009) Iwf, wikipedia and the "wayback machine". [Online]. Available: https://www.cl.cam.ac.uk/~rnc1/talks/090528-uknof13.pdf

[61] Broadcom. (2020) Rulespace. [Online]. Available: https://www.broadcom.com/company/partners/symantec/programs/oem-sales-program/rulespace

[62] D. Fifield, N. Hardison, J. Ellithorpe, E. Stark, D. Boneh, R. Dingledine, and P. Porras, "Evading censorship with browser-based proxies," in *International Symposium on Privacy Enhancing Technologies Symposium*. Springer, 2012, pp. 239–258.

[63] A. Mathrani and M. Alipour, "Website blocking across ten countries: A snapshot." in *PACIS*, 2010, p. 152.

[64] I. Society. (2017) Internet society perspectives on internet content blocking: An overview. [Online]. Available: https://www.internetsociety.org/resources/doc/2017/internet-content-blocking/

[65] J. Polpinij, C. Sibunruang, S. Paungpronpitag, R. Chamchong, and A. Chotthanom, "A web pornography patrol system by content-based analysis: In particular text and image," in *2008 IEEE International Conference on Systems, Man and Cybernetics*. IEEE, 2008, pp. 500–505.

[66] H. Ma, "Fast blocking of undesirable web pages on client pc by discriminating url using neural networks," *Expert Systems with Applications*, vol. 34, no. 2, pp. 1533–1540, 2008.

[67] I. Dubrawsky, "Firewall evolution-deep packet inspection," *Security Focus*, vol. 29, 2003.

[68] C. Xu, S. Chen, J. Su, S.-M. Yiu, and L. C. Hui, "A survey on regular expression matching for deep packet inspection: Applications, algorithms, and hardware platforms," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 4, pp. 2991–3029, 2016.

[69] M. Broersma. (2019) Privacy groups warn on isps' 'widespread' use of deep packet inspection. [Online]. Available: https://www.silicon.co.uk/workspace/privacy-group-warns-of-isps-widespread-use-of-deep-packet-inspection-255505

[70] C. Cimpanu. (2019) 33across. [Online]. Available: https://www.zdnet.com/article/186-eu-isps-use-deep-packet-inspection-to-shape-traffic-break-net-neutrality/

[71] M. Dornseif, "Government mandated blocking of foreign web content," *arXiv preprint cs/0404005*, 2004.

[72] erdgeist. (2014) Chaos computer club on the blocking of our website in uk. [Online]. Available: https://www.ccc.de/en/updates/2014/ccc-censored-in-uk

[73] Broadcom. (2019) Financial news release. [Online]. Available: https://investors.broadcom.com/news-releases/news-release-details/broadcom-acquire-symantec-enterprise-security-business-107

[74] S. Ghosh. (2014) Web blocking: 12 symantec staff review 2,000 sites a day. [Online]. Available: https://web.archive.org/web/20150520100943/https://pcpro.co.uk/news/broadband/386965/web-blocking-12-symantec-staff-review-2-000-sites-a-day/

[75] B. telecom. Bt wi-fi protect infopaper.

[76] M. S. Schmidt, K. Bradsher, and C. Hauser, "Us panel cites risks in chinese equipment," *The New York Times*, vol. 8, 2012.

[77] D. Lee. (2013) David cameron: "in the balance between freedom and responsibility we have neglected our responsibility to children". [Online]. Available: https://www.bbc.com/news/technology-23452097

[78] ORG. (2020) Blocked org webpage. [Online]. Available: https://www.blocked.org.uk/

[79] ——. (2020) Blocked org - run your probe. [Online]. Available: https://www.blocked.org.uk/run-your-own-probe

[80] ——. (2020) Github blocking rules. [Online]. Available: https://github.com/openrightsgroup/Blocking-Middleware/tree/master/config

[81] AAISP. (2020) Why are we against censorship? [Online]. Available: https://www.aa.net.uk/broadband/real-internet/

[82] A. Filasto and J. Appelbaum, "Ooni: Open observatory of network interference." in *FOCI*, 2012.

[83] OONI. (2020) Ooni probe tests. [Online]. Available: https://ooni.org/nettest/

[84] ——. (2020) Github ooni web connectivity specification. [Online]. Available: https://github.com/ooni/spec/blob/master/nettests/ts-017-web-connectivity.md

[85] C. Lab and Others, "Url testing lists intended for discovering website censorship," 2014, https://github.com/citizenlab/test-lists. [Online]. Available: https://github.com/citizenlab/test-lists

[86] OONI. (2020) Ooni api. [Online]. Available: https://api.ooni.io/

[87] GGPview. (2020) Ggpview webpage. [Online]. Available: https://www.bgpview.io/

[88] dbip. (2020) dbip webpage. [Online]. Available: https://db-ip.com/

[89] O2. (2020) O2 age verification page. [Online]. Available: https://shieldav.o2.co.uk/

[90] OONI. (2020) Ooni export full dataset. [Online]. Available: https://ooni.org/post/mining-ooni-data

[91] R. Mitchell, *Web scraping with Python: Collecting more data from the modern web*. " O'Reilly Media, Inc.", 2018.

[92] J. Services. (2020) Jisc services website. [Online]. Available: https://www.jisc.ac.uk/

[93] D. Communication. (2020) Daisy communication website. [Online]. Available: https://daisycomms.co.uk/

[94] BT. (2020) Bt strict parental control. [Online]. Available: https://www.bt.com/help/security/parental-controls/how-to-keep-your-family-safe-online-with-bt-parental-controls-an

[95] ORG. (2020) Org blocked faqs. [Online]. Available: https://www.blocked.org.uk/faqs#https

[96] R. Chirgwin. (2020) Ietf protects privacy and helps net neutrality with dns over https. [Online]. Available: https://www.theregister.co.uk/2017/12/14/protecting_dns_privacy/

[97] M. P. Van der Loo, "The stringdist package for approximate string matching," *The R Journal*, vol. 6, no. 1, pp. 111–122, 2014.

[98] R. A. Wagner and R. Lowrance, "An extension of the string-to-string correction problem," *Journal of the ACM (JACM)*, vol. 22, no. 2, pp. 177–183, 1975.

[99] W. Härdle and L. Simar, *Applied multivariate statistical analysis*, 2012.

[100] K. Backhaus, B. Erichson, W. Plinke, and R. Weiber, *Multivariate Analysemethoden*. Springer, 2016.

[101] Wikipedia. Shotacon. [Online]. Available: https://en.wikipedia.org/wiki/Shotacon

[102] Savedotorg. (2019) stop the sale. [Online]. Available: https://savedotorg.org/

[103] M. Jackson. (2015) 20k council homes in wandsworth to get 1000mbps fibre broadband - ispreview uk. [Online]. Available: https://www.ispreview.co.uk/index.php/2015/09/battersea-council-estate-homes-hail-1086mbps-fibre-broadband-pilot.html

[104] A. Ferguson. (2017) Sky q to allow people to do away with satellite dish — thinkbroadband. [Online]. Available: https://www.thinkbroadband.com/news/7623-sky-q-to-allow-people-to-do-away-with-satellite-dish

[105] M. E. Peters and D. Lecocq, "Content extraction using diverse feature sets," in *Proceedings of the 22Nd International Conference on World Wide Web*, 2013, pp. 89–90.

Table XII
MEASUREMENTS OF UNENCRYPTED WEB CONTENT (HTTP PROTOCOL) FROM THE OONI MEASUREMENTS

ASN Network	**Total Measurements**	**Detected Blocks**	**% of Blocks**	**Distribution of Blocking Techniques**			
				dns	**http-failure**	**http-diff**	**tcp_ip**
Daisy Communications Ltd	24,703	479	1.94%	74	0	405	0
Jisc Services Limited	12,783	257	2.01%	41	40	158	18
KubeNET	4,599	252	5.48%	35	7	209	1
BTnet UK Regional network	2,992	278	9.29%	93	1	178	6
M247 Ltd	2,385	173	7.25%	104	42	18	9
Virgin Media	2,163	212	9.80%	61	83	68	0
UK Internet Service Provider	1,548	96	6.20%	45	8	36	7
Sky UK Limited	1,463	925	63.23%	127	5	792	1
Vodafone Limited	1,196	75	6.27%	25	5	41	4
ONLINE S.A.S.	1,120	696	62.14%	140	548	4	4
Hutchison 3G UK Limited	1,112	149	13.40%	73	1	75	0
Telefonica UK Limited	927	375	40.45%	1	2	371	1
EE Ltd	888	31	3.49%	0	1	30	0
TalkTalk Communications Limited	758	80	10.55%	3	0	77	0
Telstra Europe Ltd	454	16	3.52%	16	0	0	0
Andrews & Arnold Ltd	358	20	5.59%	4	0	16	0
Hydra Communications Ltd	250	22	8.80%	6	0	16	0
TalkTalk Broadband	212	16	7.55%	2	1	9	4
Datacamp Limited	197	20	10.15%	0	2	18	0
VData Ltd	191	37	19.37%	0	0	37	0

Table XIII
MEASUREMENTS OF ENCRYPTED WEB CONTENT (HTTP PROTOCOL) FROM THE OONI MEASUREMENTS

ASN Network	**Total Measurements**	**Detected Blocks**	**% of Blocked**	**Distribution of Blocking Techniques**			
				dns	**http-failure**	**http-diff**	**tcp_ip**
Daisy Communications Ltd	13,793	5	0.04%	1	2	1	1
Jisc Services Limited	4,789	44	0.92%	7	37	0	0
BTnet UK Regional network	1,910	78	4.08%	33	45	0	0
KubeNET	1,642	52	3.17%	3	49	0	0
Virgin Media	1,398	92	6.58%	85	3	4	0
M247 Ltd	908	46	5.07%	22	20	0	4
UK Internet Service Provider	881	31	3.52%	22	7	0	2
Vodafone Limited	694	20	2.88%	12	8	0	0
Hutchison 3G UK Limited	650	22	3.38%	0	22	0	0
EE Ltd	571	38	6.65%	2	36	0	0
ONLINE S.A.S.	523	256	48.95%	80	176	0	0
Sky UK Limited	466	4	0.86%	0	4	0	0
TalkTalk Communications Limited	465	5	1.08%	0	5	0	0
Telefonica UK Limited	460	58	12.61%	2	56	0	0
Andrews & Arnold Ltd	215	0	0.00%	0	0	0	0
Telstra Europe Ltd	189	0	0.00%	0	0	0	0
TalkTalk Broadband	160	2	1.25%	1	1	0	0
Hydra Communications Ltd	156	0	0.00%	0	0	0	0
OVH SAS	149	10	6.71%	0	2	8	0
Datacamp Limited	134	3	2.24%	0	3	0	0
VData Ltd	130	5	3.85%	0	5	0	0

Table XIV
ISPs AND THEIR SPECIFIC CONFIGURATIONS (E.G. STRICT BLOCKING, UNFILTERED) FOR WHICH PROBES ARE SET UP BY THE BLOCKED ORG PROJECT.

ISP	First Measurement	Latest Measurement	Blocking Rules Since	Tested Sites	Detected Blocked Sites	% Blocked
AAISP	2014-05-24	2020-01-16	-	6,139,914	0	0.00%
BT	2014-07-03	2020-01-17	2014-04-23	6,262,269	178,444	2.85%
BT-Strict	2014-06-03	2020-01-17	2017-08-24	2,613,434	277,211	10.61%
BT-Unfiltered	2018-01-22	2020-01-16	2018-01-22	3,871,155	703	0.02%
EE	2014-07-02	2020-01-17	2014-05-21	3,433,894	127,264	3.71%
Gigaclear	2017-12-21	2020-01-16	-	2,072,906	0	0.00%
IFNL	2018-03-16	2018-12-01	-	810,027	0	0.00%
Kingston Communications	2017-11-21	2020-01-16	-	2,072,454	0	0.00%
LCHost	2018-01-03	2019-12-18	-	2,068,557	0	0.00%
O2	2014-05-24	2020-01-17	2014-05-18	3,642,940	126,859	3.48%
OpenDNS	2018-07-15	2018-12-01	2018-07-16	481,372	8,453	1.76%
Plusnet	2014-07-02	2020-01-17	2014-04-30	5,966,731	47,305	0.79%
Plusnet-unfiltered	2018-01-08	2018-07-15	2018-01-14	1,812,968	575	0.03%
Sky	2014-07-02	2020-01-17	2014-06-03	6,160,023	159,926	2.60%
Sky-unfiltered	2018-01-08	2020-01-16	2018-01-14	2,949,114	862	0.03%
TalkTalk	2014-07-03	2020-01-17	2014-05-01	5,544,256	201,269	3.63%
TalkTalk-unfiltered	2018-01-09	2020-01-16	2018-01-14	3,873,765	787	0.02%
Three	2014-05-04	2020-01-17	2014-05-04	2,630,665	120,309	4.57%
Uno	2016-12-05	2020-01-16	-	1,512,812	0	0.00%
VirginMedia	2014-07-02	2020-01-17	2014-06-03	5,698,031	149,298	2.62%
VirginMedia-unfiltered	2018-01-14	2020-01-16	2018-01-14	3,871,377	869	0.02%
Vodafone	2014-05-24	2020-01-17	2014-04-19	4,529,449	110,577	2.44%
Zen Internet	2017-12-21	2020-01-16	-	2,072,115	0	0.00%

Table XV
30 MOST FREQUENT WORDS FOR THE VARIABLES DESCRIPTION, TITLE AND KEYWORDS FOR THE CATEGORIES PORN AND DRUGS

Rank	**Porn**						**Drugs**					
	Description		**Title**		**Keywords**		**Description**		**Title**		**Keywords**	
	Word	***Freq***	***Word***	***Freq***	***Word***	***Freq***	***Word***	***Freq***	***Word***	***Freq***	***Word***	***Freq***
1	porn	973	com	1186	porn	760	cannabis	96	cannabis	109	cannabis	73
2	free	871	porn	1008	sex	749	online	67	com	67	marijuana	55
3	videos	833	sex	750	free	545	marijuana	64	marijuana	63	drug	54
4	sex	813	free	745	videos	518	information	50	home	47	drugs	34
5	com	664	videos	662	movies	289	drug	41	cbd	44	weed	29
6	best	612	tube	374	girls	279	best	39	online	40	hemp	27
7	looking	435	xxx	328	xxx	270	buy	38	buy	33	medical	27
8	find	396	movies	288	tube	245	cbd	38	headshop	30	online	25
9	girls	388	girls	263	adult	244	products	35	shop	29	psychedelic	20
10	movies	356	porno	249	video	234	shop	34	drug	29	cbd	20
11	site	346	information	227	gay	195	free	31	hemp	28	buy	19
12	xxx	346	resources	224	teen	195	medical	30	oil	23	psychoactive	19
13	tube	294	hot	186	porno	189	hemp	30	sale	23	headshop	18
14	hot	284	adult	183	nude	186	quality	27	seeds	22	high	17
15	watch	279	gay	174	amateur	180	including	25	best	21	shop	16
16	first	273	sale	170	hot	177	und	25	information	21	marihuana	16
17	adult	266	video	168	com	175	high	24	erowid	21	prohibition	15
18	source	264	website	167	sexy	157	headshop	23	pilze	21	mushrooms	15
19	information	262	loading	159	pictures	150	com	23	pilzgalerie	21	seeds	15
20	online	259	net	159	pics	142	find	23	addiction	17	best	14
21	sexy	259	amateur	159	chat	140	can	22	medical	15	smoking	14
22	porno	244	live	157	naked	134	grow	21	vault	15	legal	14
23	hope	240	site	156	big	128	seeds	21	org	13	addiction	14
24	expect	239	best	153	hardcore	122	get	19	weed	13	alcohol	14
25	general	239	teen	152	live	121	oil	19	news	13	law	14
26	searching	239	chat	151	asian	121	world	18	vaporizer	12	entheogen	14
27	topics	239	online	148	pussy	121	sale	18	grow	12	lsd	14
28	sale	232	sexy	142	mature	120	links	18	legal	12	thc	14
29	video	215	shop	137	anal	116	legal	18	www	12	grow	14
30	see	212	nude	130	webcam	114	order	17	products	12	oil	14

Table XVI
30 MOST FREQUENT WORDS FOR THE VARIABLES DESCRIPTION, TITLE AND KEYWORDS FOR THE CATEGORIES DATING AND WEAPONS

Rank	Dating						Weapons					
	Description		Title		Keywords		Description		Title		Keywords	
	Word	*Freq*	*Word*	*Freq*	*Word*	*Freq*	*Word*	*Freq*	*Word*	*Freq*	*Word*	*Freq*
1	dating	1140	dating	1202	dating	1085	gun	186	home	262	gun	196
2	site	725	com	591	singles	719	und	166	com	236	guns	154
3	free	691	singles	560	online	533	online	137	gun	167	rifle	129
4	singles	627	online	435	chat	399	firearms	137	schuetzenverein	145	hunting	115
5	online	613	site	422	free	367	accessories	136	firearms	103	shooting	109
6	find	503	free	296	personals	359	shooting	121	guns	99	firearms	107
7	meet	414	chat	233	site	349	shop	112	online	88	knives	97
8	love	360	meet	212	love	344	guns	107	knives	88	custom	97
9	com	307	women	161	single	310	best	106	shop	86	rifles	84
10	single	296	love	152	women	292	quality	102	sale	80	pistol	81
11	chat	295	single	150	date	266	com	99	shooting	72	knife	75
12	women	280	date	130	meet	265	custom	97	und	71	ammunition	75
13	join	272	find	125	marriage	245	hunting	96	accessories	70	tactical	74
14	looking	241	service	110	service	200	knives	94	welcome	65	accessories	73
15	men	217	gay	106	girls	187	der	93	custom	58	revolver	69
16	date	207	matrimony	104	men	172	world	90	inc	57	schuetzenverein	68
17	today	206	personals	103	romance	160	products	90	page	55	shotgun	66
18	people	205	marriage	103	sites	160	rifles	83	der	54	online	63
19	best	204	best	101	match	145	sale	82	startseite	53	arms	62
20	now	185	rencontre	98	matchmaking	145	fuer	82	rifles	49	waffen	61
21	service	177	home	96	find	143	one	70	waffen	48	ammo	60
22	und	152	partnersuche	92	gay	125	buy	69	news	46	shop	60
23	world	145	sites	91	matrimonial	122	rifle	68	rifle	45	schiessen	56
24	marriage	141	men	82	brides	121	des	67	ammunition	44	sport	51
25	website	138	matchmaking	80	com	114	site	66	arms	42	colt	51
26	new	135	russian	79	woman	112	schuetzenverein	65	www	42	ruger	47
27	sites	124	new	77	relationship	108	find	65	swords	41	shotguns	47
28	register	123	matrimonial	75	personal	104	den	64	gear	40	military	47
29	friends	120	people	71	dates	103	free	62	website	39	remington	46
30	community	118	und	70	matrimony	101	new	61	holsters	39	pistole	46

Table XVII
30 MOST FREQUENT WORDS FOR THE VARIABLES DESCRIPTION, TITLE AND KEYWORDS FOR THE CATEGORIES SUICIDE AND ANONYMIZERS

Rank	Suicide						Anonymizers					
	Description		Title		Keywords		Description		Title		Keywords	
	Word	*Freq*	*Word*	*Freq*	*Word*	*Freq*	*Word*	*Freq*	*Word*	*Freq*	*Word*	*Freq*
1	suicide	11	suicide	9	suicide	5	free	924	com	883	free	593
2	information	7	home	8	human	4	looking	820	free	631	online	453
3	life	7	exit	6	self	4	online	673	online	587	proxy	445
4	assisted	7	dying	6	euthanasia	4	best	590	proxy	411	download	411
5	site	6	information	5	society	3	find	474	information	405	web	257
6	tips	5	dignity	5	forum	3	download	464	resources	402	movies	256
7	die	4	death	4	boulimie	3	proxy	451	watch	326	unblock	252
8	end	4	end	4	assisted	3	source	447	movies	301	torrent	236
9	exit	4	com	4	death	3	links	442	net	299	watch	220
10	euthanasia	4	association	4	deliverance	3	information	435	download	273	torrents	178
11	eating	3	die	4	die	3	related	425	loading	214	anonymous	169
12	may	3	right	4	dying	3	first	417	org	198	site	166
13	leben	3	society	4	right	3	general	417	torrent	176	vpn	155
14	sterben	3	assisted	4	binge	2	see	415	website	174	music	150
15	self	3	euthanasia	4	eating	2	topics	412	sale	167	streaming	147
16	membership	3	life	3	patientenschutz	2	searching	403	web	162	games	139
17	anorexie	3	forums	3	patientenverfuegung	2	hope	402	vpn	147	series	139
18	boulimie	3	eating	3	selbstbestimmung	2	movies	401	www	138	upload	138
19	les	3	deutsche	3	sterben	2	watch	401	live	133	video	130
20	pour	3	self	3	terminal	2	expect	398	anonymous	128	share	120
21	sur	3	emotional	3	laws	2	com	383	series	124	stream	117
22	commit	3	help	3	living	2	sale	351	websites	120	websites	117
23	don	3	hope	3	life	2	website	347	torrents	113	internet	116
24	drop	3	lost	3	therapy	2	domain	263	service	109	server	116
25	drugs	3	methods	3	ana	2	web	258	unblock	107	bittorrent	109
26	everything	3	statistics	3	pro	2	unblock	251	home	100	movie	108
27	hanging	3	accueil	3	alimentaire	2	may	218	server	96	file	107
28	hope	3	anorexie	3	anorexie	2	websites	212	movie	94	software	98
29	lost	3	boulimie	3	comportement	2	can	194	nbspthis	94	live	98
30	reading	3	hemlock	3	sit	2	music	191	unblocks	93	list	97