CCCC: Corralling Cookies into Categories with CookieMonster

XUEHUI HU, King's Collage London, UK
NISHANTH SASTRY, University of Surrey, UK
MAINACK MONDAL, Indian Institute of Technology Kharagpur, India

Browser cookies are ubiquitous in the web ecosystem today. Although these cookies were initially introduced to preserve user-specific state in browsers, they have now been used for numerous other purposes, including user profiling and tracking across multiple websites. This paper sets out to understand and quantify the different uses for cookies, and in particular, the extent to which targeting and advertising, performance analytics and other uses which only serve the website and not the user add to overall cookie volumes. We start with 31 million cookies collected in Cookiepedia, which is currently the most comprehensive database of cookies on the Web. Cookiepedia provides a useful four-part categorisation of cookies into strictly necessary, performance, functionality and targeting/advertising cookies, as suggested by the UK International Chamber of Commerce. Unfortunately, we found that, Cookiepedia data can categorise less than 22% of the cookies used by Alexa Top20K websites and less than 15% of the cookies set in the browsers of a set of real users. These results point to an acute problem with the coverage of current cookie categorisation techniques.

Consequently, we developed *CookieMonster*, a novel machine learning-driven framework which can categorise a cookie into one of the aforementioned four categories with more than 94% F1 score and less than 1.5 ms latency. We demonstrate the utility of our framework by classifying cookies in the wild. Our investigation revealed that in Alexa Top20K websites necessary and functional cookies constitute only 13.05% and 9.52% of all cookies respectively. We also apply our framework to quantify the effectiveness of tracking countermeasures such as privacy legislation and ad blockers. Our results identify a way to significantly improve coverage of cookies classification today as well as identify new patterns in the usage of cookies in the wild.

CCS Concepts: • **Security and privacy** → *Web application security*; • **Computing methodologies** → *Machine learning approaches.*

Additional Key Words and Phrases: Web tracking, Third-Party Cookie, Cookie Categorisation

ACM Reference Format:
Xuehui Hu, Nishanth Sastry, and Mainack Mondal. 2021. CCCC: Corralling Cookies into Categories with *CookieMonster*. In *13th ACM Web Science Conference 2021 (WebSci '21), June 21–25, 2021, Virtual Event, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3447535.3462509

1 INTRODUCTION

First introduced in the mid-nineties as a way of recording client-side state [Netscape 2002], cookies have proliferated widely on the Web, and have become a fundamental part of the Web ecosystem. However, there is widespread concern that cookies are being abused to track and profile individuals online for commercial, analytical and various other purposes [Sanchez-Rola et al. 2019]. Recently, there has been a movement to restrict their usage, and companies such as Google have announced plans to replace certain kinds of cookies with more privacy-friendly equivalents [Bindra 2021]. Before such drastic changes, however, it is important to take stock and understand how cookies are really being used across the Web. Given the variety and number of uses for cookies and the fact that practically every website uses them, this is a herculean task.

This paper is a first attempt to address this problem and catalogue cookies in-the-wild. Currently the most commonly used classification in English language websites is the one proposed by the UK International Chamber of Commerce (UK ICC). The UK ICC catalogues cookies into four broad categories [ICC 2012]: *strictly necessary* cookies, which are essential for the website's function (e.g., logins, shopping carts); *performance* cookies, which collect analytics information to improve a website's performance; *functionality* cookies which remember user choices such as preferred language or location, allowing personalisation of the website to the user; and *targeting/advertising* cookies, typically placed by third party advertising networks with the permission of the first party website to profile users and serve them ads.

Our starting point is Cookiepedia, a database of over 31 Million cookies, which are categorised into the four UK ICC categories. Unfortunately, however, our measurements show that when queried with the cookies from the Top20K websites according to Alexa[1], Cookiepedia can only identify and categorise around 22% of the cookies. We then turn to a Chrome plugin which some of us developed previously [Hu et al. 2020], and is currently being used by over six thousand users. 475 of these users (from 44 countries) are continuously donating anonymised cookie data to us[2]. Cookiepedia coverage on this dataset is even lower – it is able to classify less than 15% of this sample of cookies in-the-wild.

To address this problem, we treat the Cookiepedia data as a giant labelled dataset of cookie categories, using which we train a number of standard machine learning models, using a standard 5-fold cross-validation. Several of these models perform well, and we obtain a best-of-class F1 measure of around 0.95 with the Multinomial Naive Bayes classifier. All our models rely on lexical n-gram features generated from the *names* of cookies. We then show that our model, which we term as *CookieMonster*, not only performs well in automatically categorising cookies found in the Cookiepedia data, but also generalises to other cookies in-the-wild. We manually classify cookies on a random selection of Alexa Top 1 Million websites that are not in Cookiepedia, by leveraging GDPR consent managers used on these websites to allow users in the EU to decline particular categories of cookies. We demonstrate that our model

WebSci '21, June 21–25, 2021, Virtual Event, United Kingdom

ACM ISBN 978-1-4503-8330-1/21/06…$15.00
https://doi.org/10.1145/3447535.3462509

[1] https://alexa.com, which provides widely used ranks for websites
[2] This study is approved by our university ethics No. MRSP-19/20-18077

is able to correctly predict (94% accuracy) the cookies which will be removed when a given category of cookies is declined through GDPR consent management, which indicates that our models are able to correctly categorise cookies in-the-wild.

Inspired by this performance on websites not represented in Cookiepedia, we then use our model on all cookies in Alexa Top20K websites, and find that the necessary and functional cookies (which are the two categories that are directly beneficial mainly to the user and not the website) constitute only 26.52% and 9.52% respectively of all cookies. Furthermore, we demonstrate for the first time that there are a number of third party cookies which are multi category. We then look at cookies donated by the users of our browser plugins, and find that even smaller percentages – less than 9.52% (respectively 13.05%) of cookies found in-the-wild are necessary (resply. functional). Interestingly, tracking/advertising cookies comprise 59.99% of cookies in the browsers of users from EU countries and a nearly similar 61.33% of cookies in non-EU countries, which is disturbing as it implies that EU users are not effectively utilising GDPR consent management to decrease the numbers of trackers in their browsers. We find similar results for other jurisdictions where there are web privacy-related laws, such as California (CCPA) or Brazil (LGPD). We also find that ad blockers are not fully effective, managing to block between 40–80% of all the third party advertising cookies.

The rest of this paper is organised as follows: section 2 presents related work. section 3 presents the design of our model, *CookieMonster*. section 4 then uses this model in-the-wild, beyond the Cookiepedia data it is trained on, and counts the numbers of cookies that are not necessary and functional and can be eliminated, and quantifies the effects of ad blockers and privacy regulations. section 5 concludes by discussing implications for future work.

2 RELATED WORK

We divide the related work in two dimensions—prior work on understanding usage of cookies with a focus on third-party cookies and prior work on categorising cookies, which tried to bring transparency into the tracking ecosystem.

Detecting third-party cookie usage in online tracking: Cookies are an integral part of the Web, and were designed to store and *remember* information across sessions about a particular user visiting a particular site[Netscape 2002]. However, cookies today are often leveraged for tracking users across services. These tracking cookies, often set by third-parties, store and commercialize information regarding browsing habits of users, often without user consent.

In fact, privacy violation by these third-party cookies has become a common problem today, e.g., while browsing news [Agarwal et al. 2020] or processing online payment [Preibusch et al. 2016] these cookies are generally placed to trace and speculate on users' online activities at scale. Consequently, a flurry of recent studies attempted to identify and detect these third-party cookies in websites. Many of these studies leverage third-party domain names in cookies to detect third party cookies [Shi et al. 2018; Zhao et al. 2019]. A few studies also leverage the similarity of source HTML codes of a website [Jain and Gupta 2018] to identify third-party presence and alert users. However, these methods are often computationally expensive and greatly affects the practicality of real-time detection [Malandrino et al. 2013]. Our study contributes to this line of study by designing *CookieMonster*, a novel machine learning-driven method for scalably categorising cookies.

Aside from academic proposals, there are a number of deployed approaches to detect third party-cookie presence and protect online users from privacy intrusion. For example, popularly used tracker-blocking lists like EasyList tried to automate detection and blocking third-party trackers. However, researchers found that EasyList can miss around 25% tracker detection [Bielova et al. 2020] and is extremely hard to be continuously updated due to ever changing lists of third party domains [Cozza et al. 2020]. Thus, our work provides a complementary machine learning-based approach for cookie categorisation and potential blocking which can be used in conjunction with these list-based approaches. In fact, our work builds on recent work that used a learning approach using web-traffic data [Kargaran et al. 2020]. This work captures invisible trackers missed by filter lists using web-traffic from user's computer and obtains 90.9% accuracy of detection for the Alexa Top10K websites. Our approach is complementary as both can be used to identify and potentially block trackers. Moreover, our system primarily depends on cookie *names* for categorization (removing the need for more computationally expensive capture and analysis of web traffic). Furthermore, we identify not only trackers, but also necessary and performance cookies and we achieve an accuracy of 94%, significantly more than prior work [Kargaran et al. 2020] for third-party tracker detection. By virtue of using cookie names, our work also evades anti-ad-blockers—tools that are being developed against ad blockers [Gupta and Panda 2020; Iqbal et al. 2017] which aims to defeat today's ad/tracker blocking systems by manipulating the webpage source code.

Categorising cookies in the online ecosystem: With the advent of General Data Protection Regulation (GDPR) in the EU, cookie categorisation has become more structured. The UK ICC has suggested a 4-part cookie categorisation which is now widely used [ICC 2012]. Cookiepdia [OneTrust 2020], a massive dataset of more than 31 million cookies collected from websites and managed by OneTrust (a company for operationalising privacy, security and data governance), classify some of their cookies into the categories suggested by ICC [Collective 2018; OneTrust 2019]. However, a recent work shows that a large number of cookies in Cookiepedia are categorised as "unknown" [Cahn et al. 2016]. Multiple studies have used Cookiepedia but completeness has been an issue, with less than 45% of cookie names being recognised [Cahn et al. 2016; Cahn et al. 2016; Urban et al. 2020], which has impacted the usability of Cookiepedia for cookie categorisation.

To that end, a few earlier studies also looked at tracker categorisation using classification techniques. For example, the timestamp or IP address embedded in cookies has been the basis of unsupervised classification of trackers [Gonzalez et al. 2017], while others use application-level traffic logs to automatically detect services running some tracking activity [Metwalley et al. 2015]. In general even more studies have attempted to detect privacy leaks via machine learning, from detecting tracking to detecting phishing [Iqbal et al. 2018; Jain and Gupta 2019; Tian et al. 2018]. In this work, we developed

CookieMonster which uses a supervised classification approach. *CookieMonster* uses Cookiepedia data as its training data to create a supervised cookie detection framework which is accurate and categorises cookies with very low latency based on features extracted for just cookie names. Furthermore, the Cookiepedia labels allow us to divide cookies into all four UK ICC categories, rather than a coarse-grained division into tracking and non-tracking cookies as in previous work.

3 COOKIEMONSTER: A SYSTEM TO UNDERSTAND COOKIE CATEGORIES

In this section, we present our attempt to categorise cookies first using Cookiepedia [OneTrust 2020] and identify its inadequacy. Then we will demonstrate how we designed *CookieMonster* using a data-driven approach to enable large-scale accurate cookie categorisation.

3.1 Inadequacy of Cookiepedia for cookie categorisation

As we mentioned in section 1, we first attempted a simple off-the-shelf approach using Cookiepedia. Cookiepedia is an open-source database of browser cookies containing cookie details as well as their categorisation according to cookie usage. Cookiepedia is maintained by OneTrust, a privacy management software company and reports existence of 31,553,377 cookies [OneTrust 2020] in their database.

Cookiepedia provides a simple online search interface to search for cookie names. To that end, we first used browser automation using Selenium [Selenium 2021] to collect all active cookies from Alexa global Top20K websites. In total these globally most popular 20,000 websites used 54,694 unique cookies (with unique cookie names, i.e., cookie identifiers) for their visitors. In order to categorise these cookies, we query Cookiepedia with each of the cookie names using a Selenium-driven automated browser. For each of these cookies, Cookiepedia returned one of six categories: Strictly Necessary Cookies (essential for features of the website), Performance Cookies (used to collect information about how visitors use a website), Functionality Cookies (allow websites to remember user preferences), Targeting / Advertising Cookies (used to deliver personalized advertisements to users), Unknown and Nonexistent. The first four of these categories are based on UK ICC categorisation, which is also used in GDPR cookie consent management platforms [Collective 2018]. An "Unknown" category indicates that the cookie exists in the Cookiepedia database but is not classified. A "Nonexistent" label indicates that a particular cookie does not exist in the Cookiepedia database.

We present the Cookiepedia-driven categorisation of 54,694 unique cookies used by 20,000 top Alexa websites (that we collected) in Table 1. We make an surprising yet important observation–43,116 (78.83%) of the cookies used by these Top20K websites simply remain uncategorised when we use Cookiepedia database. Thus, even a massive database like Cookiepedia simply fell short in categorising the majority of the cookies used in even most popular websites today. To that end, in order to improve the categorisation of cookies while ensuring high accuracy and coverage we design and evaluate *CookieMonster*.

Cookie Category	# cookies	% cookies
Strictly Necessary	3,071	5.61
Functionality	1,102	2.01
Performance	3,025	5.53
Targeting/Advertising	4,380	8.01
Unknown	19,007	34.75
Nonexistent	24,108	44.08
Unknown+Nonexistent	43115	78.83
Total	54,694	100

Table 1. Cookie categorisation using Cookiepedia for cookies used by Alexa global Top20K websites. The first four categories align with the UK ICC categories. Cookiepedia returns "nonexistent" when the cookie name does not exist in its database, and "unknown" when the cookie name exists in the database but has not been categorised. 78.83% of cookies from Alexa Top20K websites are either unknown or nonexistent.

3.2 CookieMonster Design

The key idea of our system is to use machine-learning for accurate cookie categorisation in the wild. The ground truth for our classifier is the cookies collected from Alexa 20k websites which is classified in one of the four meaningful categories via Cookiepedia. There were 11,578 such cookies with their categorisation into four categories–Strictly Necessary, Functionality, Performance, Targeting/Advertising (Table 1). For these cookies we used features extracted from the *cookie names* to train our classifier.

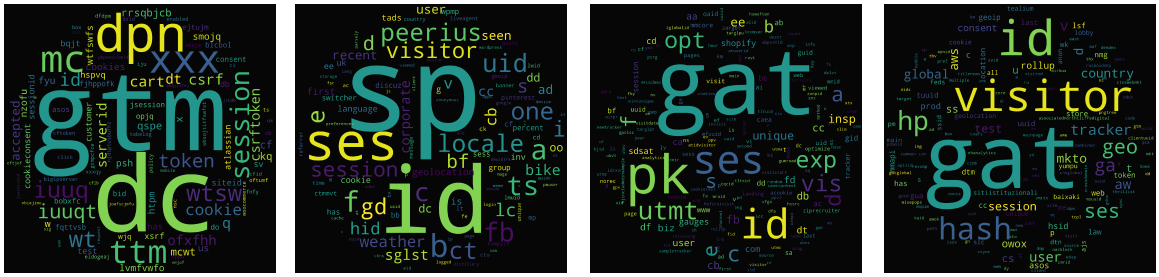
3.2.1 Preprocessing and tokenising cookie names.

Each cookie is a name-value pair and the cookie-name is unique for each cookie. We noted via manual inspection that cookie names can be meaningful and appear to provide some hints about functionality. Thus we decided to use features extracted from these names for categorisation. First, we removed all numbers from each cookie name ((e.g., ADS_324 became ADS_). Next, we tokenise these names using punctuation characters (e.g., %, ~, ., _, -). Thus, at the end of preprocessing and tokenization, a cookie with the name *gdpr-track-status45* will be split into tokens *"gdpr", "track", "status"*. Furthermore, we split the resultant token using capitalization (i.e., AnalysisUserId → [Analysis, User, Id]) and used the enchant dictionary [Thomas 2020] to segment known word combinations into root words (i.e, dayssincevisit → [days, since, visit]). Finally, we case-folded all the resulting tokens. In total, after this tokenization, we retrieved a total of 2,504 unique tokens from 11,578 cookies in our ground truth data.

3.2.2 Manually checking correlation of cookie categories and tokens.

Next, to verify the resultant tokens are meaningful, we divided the names into the four cookie categories as provided by Cookiepedia. We focused on the most popular tokens for each of our four cookie categories. We present the top tokens in each category from cookies in Figure 1 via wordclouds. To increase readability we show only tokens from top 100 domains in the figure.

We immediately notice that some particular tokens and token combinations were immensely frequent in cookies from specific cookie categories. For example, cookie combinations like (gat, gtag)



(a) Strictly Necessary (b) Functionality (c) Performance (d) Targeting/Advertising

Fig. 1. Wordclouds showing the most frequent tokens within cookie names from each cookie category.

are popular within Targeting/advertising cookie names. In fact, many popular tokens (e.g., geo, country, location, global) given by the cookie names in Targeting/advertising categories identify their usage in location tracking. Furthermore, names of third-party trackers are also frequent in these tokens (e.g., OWOX, Marketo, demdex). Although preliminary, our manual inspection of tokens gives us confidence that these tokens are correlated with cookie categories and using them as features in a supervised learning framework has the potential to be successful.

3.3 Supervised Cookie Categorisation in CookieMonster

3.3.1 Training a classifier for CookieMonster. We model the cookie categorisation as a supervised multi-class classification problem to predict our four cookie categories—strictly necessary, functionality, performance and targeting/advertising. Given a cookie name, we extracted the tokens from the names (as mentioned above) and used them as features. Consequently, we evaluated seven classification algorithms to check the performance and identify which one to use in *CookieMonster*. We used the known categorises of cookies (from cookipedia) as our training data. Specifically, we evaluated Multinomial Naive Bayes (MNB), Softmax Regression (Multi-layer perceptron or MLP), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest, Naive Bayes and Binary Search Tree (BST). We used a 5-fold cross validation with 80-20 split between training and testing data. We used overall (Micro) precision, recall and F1-score over all-classes to report the accuracy of categorisation for all of the seven models in Table 3.

We make two observations from this table: First, the top four algorithms according to F1-score (MNB, MLP, SVM, KNN) all achieved F1-scores more than 0.9, signifying the utility of our proposed features based on tokenising cookie names. Second, the top two algorithms (MNB and MLP) both achieved a F1-score of more than 0.94, making them suitable for use in *CookieMonster*. To that end, given we envision *CookieMonster* to be used in the wild for cookie categorisation, we next check the average categorisation latency for all of these classifiers.

3.3.2 Latency of prediction for classifiers. We present the average prediction latency for predicting the category of a single cookie during testing in Table 3. We note that, models like Bernoulli Naive Bayes, although extremely fast, provides a relatively poor F1-score (0.83). To that end, we focused on the top two classification models (MNB and MLP). These two models, while ensuring an F1-score of nearly 0.95, are quite different in terms of prediction latency. In fact, MLP has an average prediction latency of 1.2860 ms which is 293% higher than MNB. Therefore, we choose this pre-trained Multinomial Naive Bayes (MNB) model to use in *CookieMonster*.

3.3.3 Characterizing Misclassified cookies in MNB classifier. We further did a simple analysis to understand why MNB model did misclassify a few cookies. We present the confusion matrix for MNB classifier from one fold of cross validation in Figure 2. This shows that out of 2,016 cookies (our test set in this fold), 316 cookies got misclassified. However, the majority (244 out of 316) of this misclassification can be attributed to Necessary cookies being predicted as Targeting/Advertising and Targeting/Advertising cookies predicted as Performance cookies. We hypothesize two reasons for this. First, the Targeting/Advertising cookies share similar tokens with other cookie category. Second, Necessary and Performance cookies might sometimes also act as Targeting/Advertising cookies. We leave exploring these avenues to future work.

Finally, we note that overall (in spite of some misclassification), the accuracy of this fast MNB-based model is quite high in our training set (trained over from 11,578 cookies), however it makes a basic assumption—tokens extracted from a new cookie name will be included into 2,504 tokens that came from 11,578 cookies in our dataset. Clearly, this assumption might not hold in the wild cookie categorisation and we might encounter *out-of-vocabulary* tokens, which *CookieMonster* will need to address when used in-the-wild.

3.3.4 N-gram-based additional categorisation for cookies with previously unseen tokens. New cookie names might contain tokens which are not in the list of 2,504 tokens seen in our training dataset of 11,578 cookies. Inability to categorise these cookies poses a challenge to the categorisation coverage of *CookieMonster*. This problem is common in NLP tasks which needs to deal with OOV (out-of-vocabulary) words (thus we will call unseen tokens OOV tokens). To solve this challenge we designed an additional n-gram based classification for new cookies.

In our approach, a new cookie name (e.g., *_bti*) with previously unseen tokens is simply divided into the constituent character n-grams (e.g., *_bti* can be split into bi-grams [('b', 't'), ('t', 'i')]). In our

Actual \ Predicted	Nec	Perf	Func	Ad
Nec	486	1	2	140
Perf	2	566	7	16
Func	1	4	195	22
Ad	2	104	6	762

Fig. 2. Confusion matrix of Multinomial Naive Bayes (MNB). Majority of the misclassification happened due to Targeting/Advertising cookies.

Algorithm	Precision	Recall	F1	Mean prediction Latency (ms)
Multinomial Naive Bayes (MNB)	0.951	0.940	0.9458	0.44
Softmax Regression (MLP)	0.944	0.948	0.9457	1.29
SVM	0.947	0.867	0.926	0.03
K-Nearest Neighbors (KNN)	0.929	0.907	0.916	3.23
Random Forest	0.886	0.770	0.778	9.73
Naive Bayes	0.798	0.747	0.833	0.02
Binary Search Tree (BST)	0.649	0.461	0.409	0.05

Fig. 3. Recall, Precision and F-score of for different classification models to categorise cookies. MNB and MLP achieved more than 94% average F1-score.

Cookiepedia dataset we noted that 75% of cookie names have 5 or less characters. So we choose to use n = 2, 3 and 4. Next we simply search for these n-grams within the set of our 11,578 cookie names and create a set of existing cookie names that contain these n-grams (e.g., *NSC_mc-vsmibti* and *gati_abtc* which matched bigram of *_bti*). Finally, out of these existing cookie names we choose the one with the least edit distance with the new cookie name and output the category of that existing cookie as predicted category of the new cookie. In our example, since edit_distance (_bti, NSC_mc-vsmibti) = 10 and edit_distance (_bti, gati_abtc) = 6, so we predict category of *_bti* to be the same as the category of *gati_abtc*.

3.3.5 Final workflow of CookieMonster. So, to summarize, *CookieMonster* used cookie names to categorise cookies. On encountering a cookie name, *CookieMonster* will run the pre-processing step and identify tokens from the cookie names. If those tokens exist in the MNB-based pretrained model, then *CookieMonster* will output the prediction of MNB classifier. Otherwise, it will use the ngram based additional classifier to find a previously seen token that is lexically similar to the new unseen token, and will predict the cookie category based on the known tokens. However, one obvious question is: *since CookieMonster primarily uses the Cookiepedia data for its design, can it accurately classify cookies in-the-wild on websites not catalogued in Cookiepedia?* We answer this question affirmatively in the next section.

4 COOKIE CATEGORISATION IN-THE-WILD

CookieMonster gives us a tool to examine a collection of cookies and categorise them into the 4 widely used UK ICC categories. We first perform a manual verification (§4.1) on websites *not* included in Cookiepedia, to show that *CookieMonster* generalises widely. Then, given that we have a reasonably accurate method to classify cookies beyond the dataset it is trained and tested on, we ask what proportion of cookies are superfluous to a user's experience of websites, looking both at the Top20K websites according to Alexa, and at cookies found in browsers of real users in-the-wild (§4.2). Finally, we use *CookieMonster* to quantify the effectiveness of current web privacy measures (§4.3).

4.1 Does CookieMonster work in-the-wild? – a manual verification

section 3 demonstrated that cookie names can reveal the purpose and UK ICC category of the cookies. While this was rigorously tested using 5-fold cross validation on Cookiepedia data, we still need to validate whether the model can correctly identify the purpose of cookies on websites which have *not* been catalogued on Cookiepedia. This is not straightforward, as the *purpose* of cookies on most websites may not be apparent.

To answer this question, we take advantage of GDPR, which holds in the European Union (and in our UK vantage point). GDPR requires websites to obtain user consent before collecting data about them. Because of this, it is extremely common to see websites using consent management banners such as the example shown in Figure 4. As in the figure, many websites use the UK ICC categories for allowing users to control their consents. Thus, a careful user can control which categories of cookies are allowed from a given website. With the website in Figure 4, users *have* to allow necessary cookies (there is no choice), but may choose to allow additional categories of cookies. For example, one user may decide to allow necessary and functional cookies. Another user may allow necessary and performance cookies instead. Clearly other combinations are also possible, including allowing three or all four categories of cookies. This is a common pattern for consent management in many websites.

We can therefore determine which cookies are in the "necessary" category by visiting the website with a clean browser (after deleting all cookies and clearing the user profile) and selecting to allow only the necessary cookies. We can then clear the user's cookie and profile information again and revisit the website, this time choosing to allow necessary *and* functional cookies. The *additional* cookies installed in this second visit can be inferred to be in the "functional" category. A similar approach can be used to determine "performance" and "advertising/targeting" cookies.

The above approach is not scalable, but serves to test whether the *CookieMonster* model "works" beyond the Cookiepedia data. To this end, we select websites that satisfy two criteria: *(i)* They are *not* indexed in Cookiepedia (to test generalisability of our model).

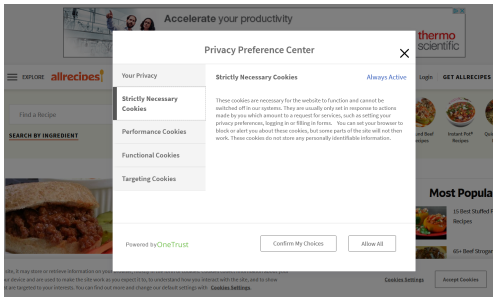


Fig. 4. Cookie Consent Example

	Recall	Precision	F1-score	OneTrust	OOV(%)
top1-100	0.93	0.87	0.91	2	0
top100-500	0.90	0.86	0.88	3	0.83%
top500-1k	0.83	0.85	0.87	0	1.68%
top1k-10k	0.86	0.93	0.89	2	0
top10k-100k	0.79	0.77	0.78	0	4.61%
top100k-1M	0.74	0.84	0.79	0	6.17%

Fig. 5. Recall, Precision and F1-score of CookieMonster for cookie recognition across Alexa top-1M websites. OOV is the percentage of cookies which were not recognised and had to be classified using the OOV technique (§3.3.4). The OneTrust column identifies the number of websites in each category using OneTrust GDPR Consent Management.

(ii) They have deployed a GDPR consent management solution that allows free choice among the four UK ICC categories (so that our approach above can be applied on that site). We randomly select $n = 60$ websites satisfying our criteria, choosing 10 each from the Alexa 1-100, 101-500, 500-1000, 1K-10K,, 10K-100K and 100K-1M ranks. We note that much of the Cookiepedia data comes from a database maintained by OneTrust[3]. Among the 60 sites we choose, 7 sites do use OneTrust (Figure 5), although these sites are still not indexed in Cookiepedia. Thus, our manual test verifies generalisability beyond Cookiepedia data to sites with and without OneTrust support.

Figure 5 shows that our model generalises extremely well. As may be expected, the performance is best for the top ranked Alexa sites (F1 score > 0.85 for the Top10K sites), but even in less popular sites up to Alexa rank 1 Million, an F1-score of > 0.78 is obtained. For each category of ranks, we also show the proportion of cookies whose names contained previously unseen tokens and therefore required the OOV technique (§3.3.4) to be used. Most cookies are recognised within our model and OOV matching is required for less than 6-7% or fewer cookies.

We conjecture that *CookieMonster* generalises beyond the Cookiepedia data it is trained on because it is based on cookie *names*, which are set by the javascript libraries or the third party providers a website uses for targeting, advertising, analytics etc. The choice of a website to use a particular GDPR consent management platform such as OneTrust (which impacts inclusion in the Cookiepedia database) is orthogonal to the libraries and third party providers (and therefore the cookie names) it uses. A few libraries and third party providers dominate the ecosystem in each country [Hu et al. 2020]; thus cookie names or the naming pattern n-grams used in *CookieMonster* generalise across websites.

4.2 What proportion of cookies are actually required for websites to function properly?

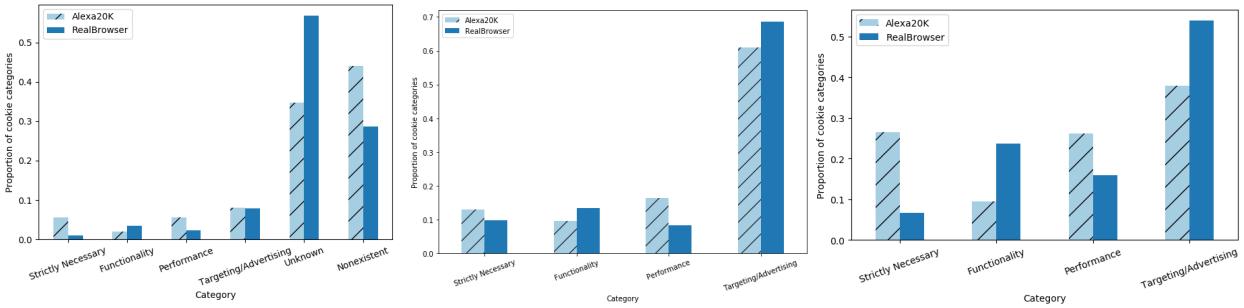
Strictly speaking, a user only needs to enable "necessary" cookies (e.g., login or shopping cart cookies). Some may choose enable "functionality" cookies that personalise a site (e.g., to user's preferred language or site layout). Arguably, performance analytics and advertising/targeting cookies benefit the website more than they do the user and do not need to be enabled. *CookieMonster* therefore provides a convenient way to quantify how many cookies are superfluous.

We study this systematically in Figure 6, by categorising all the cookies of the Alexa Top20K websites as well as cookies collected from users of a browser extension we developed and deployed in an earlier study [Hu et al. 2020], and is currently being used by over 6000 users. Specifically, in this work we use 44,971 cookies collected between November 2020 to February 2021 from 475 of these users (from 44 countries) who are donating their data. We use two methods for the categorisation: looking up the cookie name in the Cookiepedia database (Figure 6(a), which presents the same information as Table 1), and using *CookieMonster* (Figure 6(b)) to predict a category. As mentioned previously (cf. subsection 3.1), the Cookiepedia database is fairly incomplete, with over 78% of cookie names either not existing in the database or not categorised; thus, for the purpose of comparing with *CookieMonster*, we replot Figure 6(a) by ignoring these unrecognised and uncategorised cookies and renormalising the remaining cookies as 100%, obtaining Figure 6(c).

Both Cookiepedia (Figures 6(a), 6(c)) as well as *CookieMonster* (Figure 6(b)) show similar trends: According to *CookieMonster*, only 13.05% of cookies are labeled as necessary, and an additional 9.52% are functional. According to Cookiepedia, 5.6% of cookies are labeled as necessary (26.52% after ignoring unrecognised/uncategorised cookies), and an additional 2.01% are functional (9.52% after ignoring unrecognised/uncategorised). Thus, both methods suggest that *the vast majority of cookies can be removed without affecting user experience.*

Interestingly, according to both *CookieMonster* (Figure 6(b)) and Cookiepedia (Figures6(a), 6(c)), real browsers have a smaller proportion of necessary cookies and more functional/targeting cookies as compared to Alexa Top20K websites. This is likely because real users' browsers have user profiles which are better established, with a browsing history and long-lived cookies that may have been set months ago, leading to better profiling and more ads/targeting cookies. In contrast, we collect cookies on Alexa Top20K websites programmatically using Selenium with a fresh user profile instance for each website, resulting in fewer ad/targeting cookies. Also, our user base is located in different countries where there may be country-specific third party trackers [Hu et al. 2020] not visible from our UK vantage point, and therefore not captured in the Alexa crawl.

[3]https://cookiepedia.co.uk/about-cookiepedia



(a) Cookiepedia (b) *CookieMonster* (c) Cookiepedia (ignoring the unrecognised and uncategorised cookie names)

Fig. 6. Proportions of different cookie categories in Alexa Top20K (shaded) and real browsers (clear), according to (a) Cookiepedia (b) CookieMonster (c) Cookiepedia (ignoring the unrecognised and uncategorised cookie names)

4.3 On the effectiveness of current web privacy measures

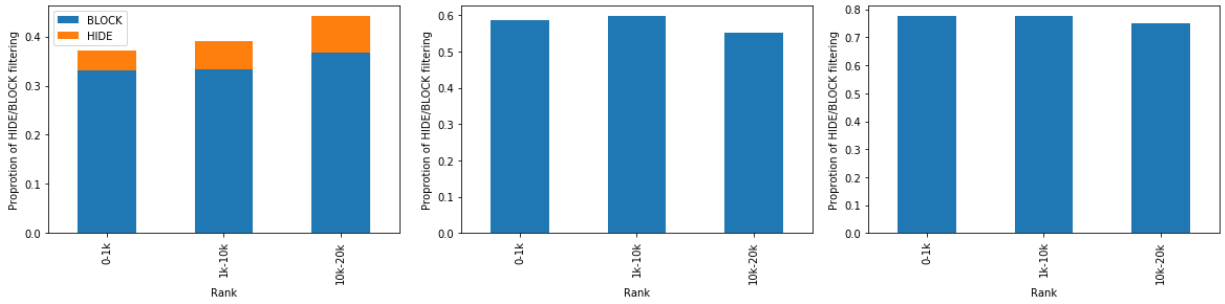
The previous section suggests that a large proportion of cookies can be eliminated from many websites without affecting their function. One of the main levers of control that users can employ to achieve this, is to use ad blockers. In addition, web privacy regulations around the world, such as GDPR, provide varying degrees of support for users to provide consent or decline different kinds of cookies. We examine their effectiveness below.

4.3.1 Ad blockers. Ad blockers typically work based on dynamically updated lists of third party advertising/targeting domains that should be blocked. Figure 7 shows how three popular block lists – EasyList, EasyPrivacy and AdGuard Plus – work on cookies found in Alexa Top20K websites. In addition to a block list, EasyList has a so-called 'hide' list of domains which break if blocked, and therefore, are loaded but not rendered on screen, to improve user experience. Unfortunately, because the domain is loaded, the user can still be tracked even if the ad itself is hidden. These domains are therefore shown separately. In general, Ad Guard appears to block a larger proportion of domains than EasyList or EasyPrivacy, even when counting cookies from hidden domains in addition to the cookies from blocked domains. We also find that there are more domains to be blocked in real browsers than when visiting Alexa Top20K sites programmatically. Again, this is likely because of additional targeting and advertising that may tend to be attracted by more mature user profiles with a continuous browsing history.

Across all the combinations tested in Figure 7, we still find that around 20% (for Ad Guard Plus) to 60% (for EasyList) of advertising and targeting-related cookies that should have been blocked are not being blocked. This is partly because the lists that ad blockers rely can never be complete. However, when we dig deeper, we on find two additional important reasons: First, ad blockers are relatively successful at blocking third party advertising and cookies, but we find that a significant proportion of first party cookies also relate to advertising. Figure 8(a) quantifies this, showing the relative proportion of targeting cookies and other categories of cookies among both first party cookies and third party cookies. Thus, several first party cookies may slip through ad blockers. Secondly, we find that both among first parties (Figure 8(c)) and third parties (Figure 8(b)), a non-trivial proportion of advertising-related domains also place other categories of cookies. Thus, a solely domain-based block list risks either blocking too much, or not covering all the domains that undertake targeting. The domain-based approach is common among all widely used ad blockers – the diversity of cookies on the web has thus far made it difficult to take a more granular approach that blocks specific cookies. However, since *CookieMonster* appears to provide reasonable predictions of cookie categories based on cookie names, we may use it as one component of a more sophisticated system that blocks specific cookies. Such approaches can complement other methods which have utilised the Internet Advertising Bureau's *Ads.txt* [Estrada-Jiménez et al. 2019] and other list-based measures to identify ads.

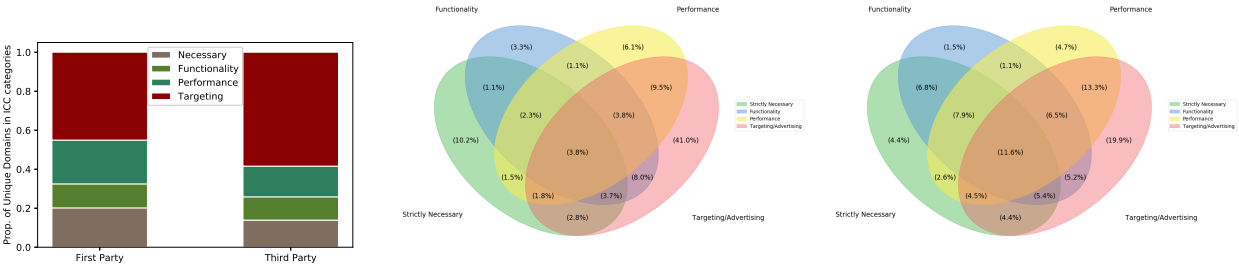
4.3.2 Privacy regulation. A second lever that users have recently obtained is support from privacy-related regulations in various legal jurisdictions. By far the most comprehensive and well-known of these is the General Data Protection Regulation (GDPR) in the EU, which introduced the notion of requiring explicit and meaningful consent. Comparable regulations include the California Consumer Privacy Act (CCPA) which allows users to opt-out of tracking and Brazil's Lei Geral de Proteção de Dados (LGPD), which is the most recent of them, and also mandates unambiguous consent from users before websites can use cookies.

Previously, using a limited cohort of 16 users, we had found that cookie numbers seen by users had not changed significantly before and after GDPR was introduced [Hu and Sastry 2019], implying that users may be choosing the 'default' choices offered by websites, which may not be privacy optimal. Here, we extend this study based on the 475 users of our extension [Hu et al. 2020] who are donating data. Specifically, we consider all users within a given privacy jurisdiction (EU, California or Brazil) and compare the proportions of ad/targeting cookies of users from within that jurisdiction to the respective proportions in browsers of users outside the jurisdiction. Figure 9 shows that in all cases, there is little difference between proportions of cookies of users within and



(a) EasyList + Alexa Topsites (b) EasyPrivacy + Alexa Topsites (c) AdGuard + Alexa Topsites

Fig. 7. EasyList, EasyPrivacy and AdGuard filter 40–80% of advertising third party cookies on Alexa Top20K sites.



(a) Relative proportions of different ICC categories in FP and TP cookies (b) Shared (TP) (c) Shared (FP)

Fig. 8. Occurrence of multipurpose Third-Party domains in Top20K websites.

out of each of the jurisdictions. This confirms (using a much larger and more representative user base) our previous finding [Hu and Sastry 2019] that users are not making the most privacy optimal choices for themselves, and may be fatigued the burden of providing consent on every website they visit, especially as several websites use dark patterns that make it difficult to choose more privacy-oriented settings [Nouwens et al. 2020].

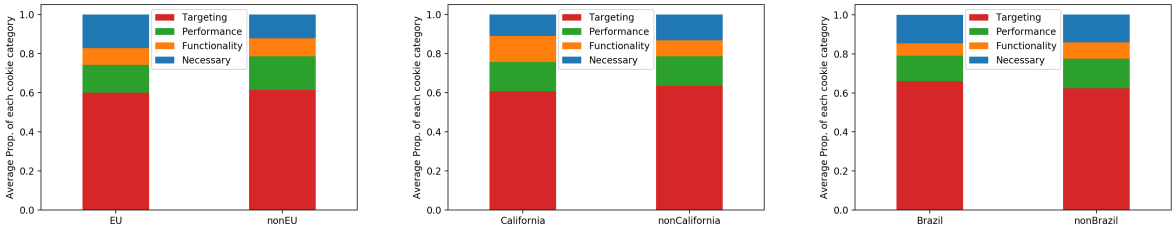
5 DISCUSSION AND CONCLUSION

This paper set out to tackle the herculean task of classifying cookies found in-the-wild. We started with data curated on Cookiepedia, and demonstrated that its coverage was inadequate – its database contained less than 22% of cookies on Alexa Top20K websites, and less than 15% of cookies found in real browsers. We therefore developed machine learning models that trained on Cookiepedia data and were also shown to work well ($F1 > 0.94$) on websites not currently in Cookiepedia. Our models use lexical features derived from cookie names, suggesting that cookie names generalise well across websites, perhaps as a result of common web templating infrastructures and libraries, and the prevalence of common third parties across websites.

We then used the trained models on Alexa Top20K websites as well as anonymised cookies donated to us by 475 users of a plugin we have developed previously [Hu et al. 2020]. We found that across the 44 countries represented in our dataset, necessary and functional cookies (the two categories beneficial to the user rather than the website) constitute only 9.79% and 13.35% of all cookies in our active countries. Thus, the vast majority of cookies can be removed without impacting website functionality or user experience.

Surprisingly we find that privacy regulations such as GDPR in the EU have not made much difference in the numbers of cookies seen by real users. This indicates that users are not effectively utilising the consent management options enabled by GDPR. Ad blockers appear to be more effective if used, but mainly focus on advertising cookies. Even among advertising cookies, a non-trivial proportion is missed because the ad blockers are based on *manually curated* lists [Atom 2005] which need to be continuously updated and because these lists are based on blocking at the level of the domains that serve up those cookies, rather than on blocking specific cookies. Unfortunately, we also find that many domains set both non-essential (e.g., advertising or performance) as well as essential (necessary or functional) cookies; thus extreme care needs to be exercised in blocking of entire domains, to ensure that functionality of the website is not broken as a result.

Thus far, the diversity of cookie names has prevented a more fine-grained approach and continuously updated but manually curated lists of domains to block have been the main tool for actively restricting tracking and cookies via ad blockers. We propose that our robust *CookieMonster* model based on lexical tokens extracted



(a) GDPR: EU vs. non-EU (b) CCPA: California vs. non-California (c) LGPD: Brazil vs. non-Brazil

Fig. 9. Proportions of ad/targeting cookies within and outside of 4 jurisdictions with privacy regulations.

from cookie names can be used as the basis for sophisticated tools enable *automatic* rejection of specific cookies belonging to categories that are not beneficial for users. We intend to develop this idea in future work.

REFERENCES

Pushkal Agarwal, Sagar Joglekar, Panagiotis Papadopoulos, Nishanth Sastry, and Nicolas Kourtellis. 2020. Stop tracking me bro! differential tracking of user demographics on hyper-partisan websites. In *Proceedings of The Web Conference 2020*. 1479–1490.

EasyList Atom. 2005. EasyList. https://easylist.to/

Nataliia Bielova, Arnaud Legout, Natasa Sarafijanovic-Djukic, et al. 2020. Missed by filter lists: Detecting unknown third-party trackers with invisible pixels. *Proceedings on Privacy Enhancing Technologies* 2020, 2 (2020), 499–518.

Chetna Bindra. 2021. Building a privacy-first future for web advertising. https://blog.google/products/ads-commerce/2021-01-privacy-sandbox/.

Aaron Cahn, Scott Alfeld, Paul Barford, and Shanmugavelayutham Muthukrishnan. 2016. An empirical study of web cookies. In *Proceedings of the 25th International Conference on World Wide Web*. 891–901.

A. Cahn, S. Alfeld, P. Barford, and S. Muthukrishnan. 2016. What's in the community cookie jar?. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 567–570.

Cookie Collective. 2018. Five Models for Cookie Law Consent.

Federico Cozza, Alfonso Guarino, Francesco Isernia, Delfina Malandrino, Antonio Rapuano, Raffaele Schiavone, and Rocco Zaccagnino. 2020. Hybrid and lightweight detection of third party tracking: Design, implementation, and evaluation. *Computer Networks* 167 (2020), 106993.

José Estrada-Jiménez, Ana Rodríguez-Hoyos, Javier Parra-Arnau, and Jordi Forné. 2019. Measuring Online Tracking and Privacy Risks on Ecuadorian Websites. In *2019 IEEE Fourth Ecuador Technical Chapters Meeting (ETCM)*. IEEE, 1–6.

Roberto Gonzalez, Lili Jiang, Mohamed Ahmed, Miriam Marciel, Ruben Cuevas, Hassan Metwalley, and Saverio Niccolini. 2017. The cookie recipe: Untangling the use of cookies in the wild. In *2017 Network Traffic Measurement and Analysis Conference (TMA)*. IEEE, 1–9.

Rohit Gupta and Rohit Panda. 2020. Block the blocker: Studying the effects of Anti Ad-blocking. *arXiv preprint arXiv:2001.09434* (2020).

Xuehui Hu, Guillermo Suarez de Tangil, and Nishanth Sastry. 2020. Multi-country Study of Third Party Trackers from Real Browser Histories. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 70–86.

Xuehui Hu and Nishanth Sastry. 2019. Characterising third party cookie usage in the eu after gdpr. In *Proceedings of the 10th ACM Conference on Web Science*. 137–141.

ICC. 2012. ICC UK Cookie guide-EU Cookie Law. https://www.cookielaw.org/media/1096/icc_uk_cookiesguide_revnov.pdf.

Umar Iqbal, Zubair Shafiq, and Zhiyun Qian. 2017. The ad wars: retrospective measurement and analysis of anti-adblock filter lists. In *Proceedings of the 2017 Internet Measurement Conference*. 171–183.

Umar Iqbal, Zubair Shafiq, Peter Snyder, Shitong Zhu, Zhiyun Qian, and Benjamin Livshits. 2018. Adgraph: A machine learning approach to automatic and effective adblocking. *arXiv preprint arXiv:1805.09155* 41 (2018).

Ankit Kumar Jain and Brij B Gupta. 2018. Towards detection of phishing websites on client-side using machine learning based approach. *Telecommunication Systems* 68, 4 (2018), 687–700.

Ankit Kumar Jain and Brij B Gupta. 2019. A machine learning based approach for phishing detection using hyperlinks information. *Journal of Ambient Intelligence and Humanized Computing* 10, 5 (2019), 2015–2028.

Amir Hossein Kargaran, Mohammad Sadegh Akhondzadeh, Mohammad Reza Heidarpour, Mohammad Hossein Manshaei, Kave Salamatian, and Masoud Nejad Sattary. 2020. On Detecting Hidden Third-Party Web Trackers with a Wide Dependency Chain Graph: A Representation Learning Approach. *arXiv preprint arXiv:2004.14826* (2020).

Delfina Malandrino, Andrea Petta, Vittorio Scarano, Luigi Serra, Raffaele Spinelli, and Balachander Krishnamurthy. 2013. Privacy awareness about information leakage: Who knows what about me?. In *Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society*. 279–284.

H. Metwalley, S. Traverso, and M. Mellia. 2015. Unsupervised Detection of Web Trackers. In *2015 IEEE Global Communications Conference (GLOBECOM)*. 1–6. https://doi.org/10.1109/GLOCOM.2015.7417499

Netscape. 2002. PERSISTENT CLIENT STATE HTTP COOKIES. https://bit.ly/3qY55Ks.

Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. 2020. Dark patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

OneTrust. 2019. OneTrust PreferenceChoice's Cookie Auto-Blocking Technology. https://bit.ly/2O4HnO6.

OneTrust. 2020. Cookiepedia. https://cookiepedia.co.uk/.

Sören Preibusch, Thomas Peetz, Gunes Acar, and Bettina Berendt. 2016. Shopping for privacy: Purchase details leaked to PayPal. *Electronic Commerce Research and Applications* 15 (2016), 52–64.

Iskander Sanchez-Rola, Matteo Dell'Amico, Platon Kotzias, Davide Balzarotti, Leyla Bilge, Pierre-Antoine Vervier, and Igor Santos. 2019. Can i opt out yet? gdpr and the global illusion of cookie control. In *Proceedings of the 2019 ACM Asia conference on computer and communications security*. 340–351.

Selenium. 2021. Selenium WebDriver. https://www.selenium.dev/

Yong Shi, Gong Chen, and Juntao Li. 2018. Malicious domain name detection based on extreme machine learning. *Neural Processing Letters* 48, 3 (2018), 1347–1357.

Reuben Thomas. 2020. Enchant. https://abiword.github.io/enchant/.

Ke Tian, Steve TK Jan, Hang Hu, Danfeng Yao, and Gang Wang. 2018. Needle in a haystack: Tracking down elite phishing domains in the wild. In *Proceedings of the Internet Measurement Conference 2018*. 429–442.

Tobias Urban, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. 2020. Beyond the Front Page: Measuring Third Party Dynamics in the Field. *arXiv preprint arXiv:2001.10248* (2020).

Hong Zhao, Zhaobin Chang, Guangbin Bao, and Xiangyan Zeng. 2019. Malicious domain names detection algorithm based on N-gram. *Journal of Computer Networks and Communications* 2019 (2019).