IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 September 2022, accepted 14 October 2022, date of publication 19 October 2022, date of current version 25 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3215753

RESEARCH ARTICLE

A Comprehensive Study of DNS Operational Issues by Mining DNS Forums

XIANRAN LIAO, JIACEN XU, (Student Member, IEEE), QIFAN ZHANG, (Student Member, IEEE), AND ZHOU LI, (Senior Member, IEEE)
Department of Electrical Engineering & Computer Science, University of California at Irvine, Irvine, CA 92697, USA

Corresponding author: Zhou Li (zhou.li@uci.edu)

This work was supported in part by U.S. NSF under Grant CNS-2047476.

ABSTRACT Domain Name System (DNS) is a fundamental component for today's Internet communications, enabling the domain-to-IP translations for billions of users and numerous applications. Yet, the operational failures of DNS are not rare and sometimes lead to severe consequences like Internet outages. To gain a better understanding of DNS operational failures, previous works examined large-scale DNS logs (DNS queries and responses between Internet users and DNS servers), but the DNS logs cannot offer a comprehensive view of the failures (e.g., errors at domain registrars) and explain the failures at a finer grain. In this paper, we try to assess DNS operational failures from another data source, the supporting forums built by DNS service providers. Specifically, we mined 4 DNS forums and crawled more than 10000 posts and 50000 replies. With a new analysis framework developed by us, we are able to tag the forum posts by different categories (e.g., general concerns, issue locations, and record types), and gain new insights regarding how and why users encounter DNS failures. In the end, we offer suggestions to DNS service providers and users to mitigate DNS operational issues.

INDEX TERMS Domain name system, forum mining, clustering, operational issues.

I. INTRODUCTION

Domain Name System (DNS) is one of the most critical Internet components, which underpins nearly every Internet activity, translating a user-friendly domain name like `www.google.com` to an IP addresses like `172.217.11.164`. A reliable DNS infrastructure is essential to the smooth operation of many Internet applications like web and email, but operational failures of DNS are not rare: a study found 13.5% DNS queries failed [1]. Moreover, DNS failures have led to severe issues like Internet outage. For example, Robinhood experienced days of service disruption because of the failures of its DNS infrastructure [2]. The IT systems such as email and Internet of the Department of Parliamentary Services in Australia were down because of DNS service failure [3].

A. PRIOR WORKS

Understanding why DNS fails in the wild is important towards improving its reliability and benefiting billions of Internet users. A number of studies have been done to inspect the DNS logs (i.e., DNS queries and responses) collected by service operators like recursive resolvers, and measure the error distributions by geo-locations, record types, etc. For instance, Yang et al. gathered a dataset with 3 billion DNS queries and found a significant higher failure ratios for AAAA and PTR records [1]. Gao et al. analyzed a dataset with 26 billion DNS query-response pairs from 600 globally distributed recursive resolvers, and found more than 50% DNS queries to the root servers failed to return the successful answers [4]. On one hand, these studies have shown important insights about DNS operational status. On the other hand, there are many questions that cannot be answered by only inspecting the DNS logs. To name a few, though a user can tell his/her DNS resolution fails from the DNS response, it is often not enough to explain *why* it fails and *how* the user should resolve (or have resolved) the issue. Beyond the

The associate editor coordinating the review of this manuscript and approving it for publication was Zhan Bu.

failures encountered by the Internet users, domain owners, website administrators, etc. also need to deal with DNS failures, but DNS logs cannot provide insights into these parties.

B. THIS STUDY

By investigating the popular DNS service providers, we found many of them have built forums to help Internet users, domain owners, website administrators, etc. debug the DNS failures. After a troubled user makes a post describing his/her encountered issue, other users or service operators/developers might post replies to help the troubled users identify the root causes and resolve the issue. Because the post and replies are usually well formatted and public, we can develop an automated framework to crawl the posts and replies related to DNS failures, classify them into different categories, measure the distribution of DNS failures by factors like network locations, and answer the aforementioned questions.

As a result, we performed a comprehensive study of DNS operational failures by mining the posts of popular DNS forums. We crawled 4 forums from Cloudflare, Comodo, OpenDNS and Spiceworks, which are either well-known public DNS resolvers or network companies, and downloaded over 10000 posts and 50000 replies in total. Then, we created a set of keyword-based filters to examine each post and assign them with different tags, e.g., issue location and record type. To derive the representative keywords, we clustered the forum posts based on k-means, and manually analyzed the representative posts of each cluster. Though there have been some studies analyzing developers' forums like stack overflow [5], [6], [7], [8], [9], [10], we have not found any study mining DNS forums. Compared to the traditional methods that analyze DNS logs passively [1], [4], our method based on forum analysis has a few advantages as listed below:

- We achieve a broader coverage of DNS failures, as forum posts can be made by any party that is involved with DNS, but DNS logs are only about DNS users' activities.
- We are able to study how DNS failures are treated after they are triggered, by analyzing the post replies.
- We could have a better understanding of the root causes of DNS failures, as DNS logs only have coarse-grained error codes (e.g., NXDOMAIN), without telling the locations and responsible parties of the errors for example.

On top of the crawled and tagged forum posts, we have gained some new insights about how DNS forums were operated, why users encountered DNS operational failures, and how the failures were solved. Here we highlight a few observations. 1) Though DNS forums have disparate activity levels from users who created posts, a post usually got multiple replies (ranging from 3.1 to 5.9 among the 4 forums), suggesting they are valid resources for a troubled user to get help. 2) Though DNS failures can be attributed to a broad range of factors (e.g., 20 record types, 9 error codes and 11 network locations were mentioned), most of the failures were actually caused by the Internet components that were on-path during DNS resolution (e.g., gateway, firewall and router) or supporting DNS functionalities (e.g., domain registrar). 3) By inspecting representative posts of each forum, we found DNS users were often troubled by the functionalities not core to DNS (e.g., Dynamic DNS provided by Cloudflare, domain blocking provided by OpenDNS and Comodo), and the problem resolution often requires the troubled user to paste the output from running the debugging tools, which could contain sensitive information. In the end, we provide a few suggestions about how to improve DNS resolution.

C. CONTRIBUTIONS

We summarize the contributions of this work below.

- We carried out the first study to understand DNS operational issues from DNS forums.
- We developed a systematic framework to mine and classify DNS forum posts.
- We conducted various measurement tasks on the data, and shed new insights into DNS operational failures.
- The source code and data of this project are released on a GitHub repository [11].

D. ROADMAP

In Section II, we overview how DNS resolution works and related works analyzing DNS operational issues, trouble tickets and forums. Section III describes how we collect and analyze the forum data. Section IV elaborates the findings we gained from different measurement tasks. Section V provides a qualitative study by analyzing prominent posts. Section VI discuss the limitations and future works, and Section VII summarizes the key findings and offers suggestions to address DNS operational issues.

II. BACKGROUND AND RELATED WORKS

In this section, we first overview the process of DNS resolution and the involved entities. Then, we review the prior studies about DNS operational issues. Finally, we describe other related works that apply content analysis.

A. DNS

The resolution of a domain name, e.g., `example.com`, can involve many entities on the network path. The client-side DNS resolver (or stub resolver), like the ones provided by operating systems and browsers, issues a request to a user-specified recursive resolver filled with the domain name, record type (e.g., A or CNAME), etc. The request can be intercepted by a DNS forwarder (e.g., integrated by a router) and served by its cache before reaching the recursive resolver. The request could either go to a public resolver (e.g., Google Public DNS) or an ISP resolver. In either case, the resolver first tries to fulfill the request with its cache. If the domain has not been resolved by any user previously or the cache has expired, the resolver will relay the request to authoritative

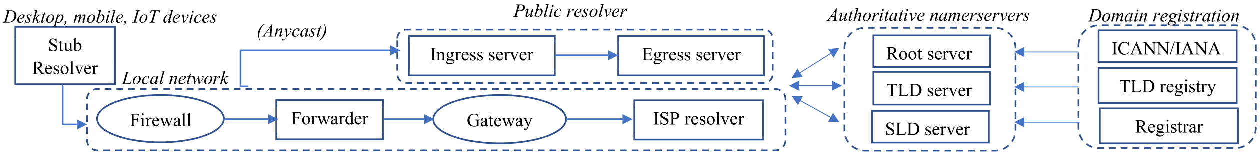


FIGURE 1. DNS infrastructure and the request flow. Boxes are clients and servers processing DNS. Circles are other entities potentially interfering DNS resolution.

nameservers, which have authentic DNS records installed by the domain owners into the zone files. The requests are handled by nameservers following the domain levels, e.g., `.` handled by the DNS root, `.com` handled by the TLD (Top-Level Domain) server, and `example.com` handled by the SLD (Second-Level Domain) server. Before a domain can be resolved, the domain owner needs to find a domain registrar, which is delegated by a TLD registry, to register the domain name and make it accessible to the general public. Figure 1 illustrates the process of DNS resolution and its connection to domain registration.

B. DNS OPERATIONAL ISSUES

Though DNS is supposed to be highly reliable due to its distributed and hierarchical structure, in reality, DNS operational issues (or failures) are not rare. In Figure 1, all network locations can introduce DNS failures, including client-side resolvers, local networks, public resolvers, authoritative nameservers, and registration services.

Previous works aim to identify the existence and explain DNS failures by analyzing the DNS traces or inspecting the configurations. Papps et al. found the configuration errors on DNS zones diminish its robustness guarantees [12], and a number of approaches have been developed to find such mis-configurations [13], [14], [15], [16], [17]. In addition to troubleshooting DNS zones, another direction is to monitor domain resolution by active probing [18], [19], [20] or passive data analysis [1], [4], [21], [22]. The measurement study by Yang et al. on 3 billion queries showed that 13.5% of them failed [1], and the failure ratios are particularly high for AAAA and PTR records (more than 50%). Lu et al. [20] leveraged peer-to-peer proxies to measure the client-side reachability and performance of public resolvers, and show that configuration issues of middleboxes and censorship severely interfered with the service quality of several public resolvers (e.g., 16% plaintext queries to Cloudflare DNS failed due to that `1.1.1.1` is reserved by devices manufactured by Cisco and AT&T, and 99.99% queries from China to Google's DNS-over-HTTPS service failed because of censorship). Besides, the mismanagement and cyber-attacks against domain registrars and registries could lead to DNS failures as well, e.g., through dangling DNS records [23] and shadowed subdomains [24].

Compared to existing works that analyze DNS operational issues, mining DNS forums achieves broader coverage of entities that are troubled by DNS failures (end-users, operators, and registrants) and gains deeper insights into the root cause and treatment of DNS failures.

C. TICKET ANALYSIS

Our work performs content analysis to measure the DNS operational issues encountered by Internet users in the wild. Similar approaches have been performed on the trouble tickets to understand the IT operational issues in general. Potharaju et al. [25] proposed NetSieve which combines statistical natural language processing (NLP), ontology modeling, etc. to automatically parse and infer the problem from tickets. Zhou et al. [26] proposed STAR to find the best resolution given a ticket summary. Many systems have been developed to improve the performance of ticket analysis, mostly facilitated by new machine-learning techniques [27], [28], [29], [30], [31], [32], [33], [34], [35], [36]. Our work performs content analysis on DNS forums, which presents different characteristics compared to tickets (e.g., less rigorous as the posts are usually written by normal Internet users rather than technicians).

D. FORUM ANALYSIS

The main data source of this measurement study is forums. Previous works have mined programming forums, e.g., stack overflow, to discover the discussion topics [5], [6], find successful answers [7], [8], their impact on software security [9], [10], etc. We mined DNS forums, which have not been extensively analyzed as far as we know. Besides forums, recent works have also mined software repositories like GitHub, and the text within commits, issues, and pull requests were analyzed to study bugs related to autonomous vehicles [37] and deep learning stacks [38].

III. DATA COLLECTION AND CLUSTERING

A. FORUM SELECTION

As the first step of this study, we select the measurement targets, i.e., DNS forums, by examining DNS services. Firstly, we surveyed existing works about DNS (e.g., [39] and [40]) and identified their studied DNS services. For each service, we searched for its DNS forum and inspected the number of posts. We consider a DNS service as a study target when its number of posts is sufficient (i.e., over 100) and it is not too outdated (i.e., the latest post/comment was observed in 2019 or after). We also searched Google to identify popular third-party forums that are not owned by any service provider.

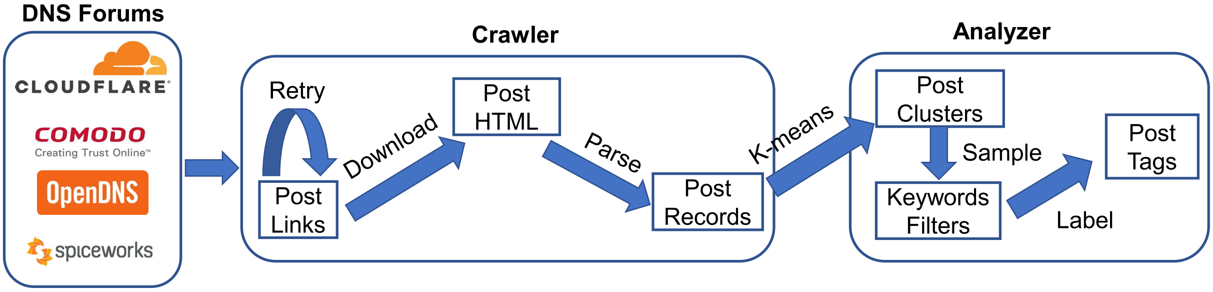


FIGURE 2. The workflow of our forum crawler and content analyzer.

In the end, we identified 4 DNS forums as the target, and below we describe each one.

- **Cloudflare** is a very popular public DNS resolver, which can be accessed by IP `1.1.1.1`. It is ranked 2nd among the public resolvers in the volume of processed DNS requests (Google Public DNS is No.1) [41]. Since Cloudflare runs a number of Internet services, like CDN, in addition to DNS, its forum posts [42] need to be filtered before the measurement study. To this end, we select the posts tagged by Cloudflare as "DNS & Network" or "1.1.1.1", and other posts containing keywords "DNS" (case insensitive).
- **Comodo** is a company that offers cloud-based cybersecurity services, and DNS resolver is included, which performs security-based filtering on the DNS requests and responses. We mined the posts under its child board "Help - DNS" [43].
- **OpenDNS** is another popular public DNS service that integrates security features, like anti-phishing protection. We examined the posts under its child board "OpenDNS Community - Community Help" [43] as the majority of the posts are about DNS.
- **Spiceworks** is a professional network technology company that runs forums to help IT staff. We found its DNS child board [44] is quite active so we include it in this study.

Besides forums, we also found code repositories (e.g., on GitHub and GitLab) of open-source DNS software containing issues about DNS, but we did not analyze them in this work, given that they are mainly facing developers rather users and they have different characteristics (e.g., the bugs are about software implementation, rather than operational issues). We discuss it further in Section VI.

B. FORUM CRAWLING

For each target forum, we firstly extract the post URLs from its entry URL. In most cases, the forum splits the posts into pages, so we can navigate to one page by changing the page number parameter in its entry URL, and then extract all post URLs. One exception is Cloudflare, which automatically loads 30 more posts when the visitor scrolls to the bottom of its webpage. After monitoring the network packet traffic, we discovered the forum automatically loaded a JSON file with the page number as a parameter to update the webpage. Therefore, instead of using Cloudflare's entry URL to get post URLs, we navigated through those JSON files.

For each post URL, we try to download its associated HTML page and retry up to three times if we encounter erroneous responses. After that, the post HTML is parsed and only relevant content will be analyzed (e.g., advertisements and banners are filtered out). Special characters and punctuation are also removed for the content analyzer.

For each crawled post, we extract 8 fields and store them in our database. The fields include the post URL, whether the post was closed, the post creator, the creation date of the post, the post content, a list of users who replied to the post, replies to the post, and a list of tags generated by the content analyzer.

Regarding the implementation details, we choose Requests version 2.25.1 to retrieve the post contents and BeautifulSoup version 4.9.3 to parse the post HTMLs. We use MongoDB version 4.2.17 to store the parsed records, and each record is stored as a JSON object. The total database storage is 3.5 GB. The crawler is deployed in a lab machine, with AMD Ryzen 9 3900X 12-Core Processor and 32 GB memory, running Ubuntu 20.04.3 LTS. Figure 2 illustrates the workflow of our crawler and content analyzer.

C. ETHICS

To avoid raising ethical issues, we follow the approaches in data collection described below. We put a rate limit on our crawler (e.g., once a crawler process finishes extracting contents from one page, it will be put to sleep for 80 seconds) and only crawl the public links allowed by robots.txt. During the period of this study, we use the IP address from our institution instead of IP addresses from proxies, so the forum admins can trace back to our crawler and notify us when they consider our data collection is intrusive. Throughout the data collection step, we have not received any complaints. Our crawler was not blocked or challenged with CAPTCHA by any forum except OpenDNS, which replied with error messages. By investigating the error message and experimenting with different crawling parameters, we found OpenDNS considered our crawling frequency too high. As a result, we extended the interval between requests from 80 seconds to 200 seconds and were able to download its posts.

D. POST CLUSTERING AND ANALYSIS

Firstly, we attempted to automatically cluster the crawled forum posts and infer their tags, like prior works [45], [46]. We examined a few clustering methods, and decided to use k-means in the end. The other methods led to worse results. For example, we tried DBScan, which needs to specify the threshold of the maximum distance and the minimum data points in a cluster, but it does not work well because a lot of data points are located distantly from each other. With k-means, we can experiment with different number of clusters k, and select the best k.

Specifically, we first pre-processed each post to lemmatize each word (based on the priority noun-adjective-verb by setting lemmatizer parameters) and removed the stop words. Then, we applied TF-IDF (Term Frequency - Inverse Document Frequency) vectorizer to convert a word sequence to a numerical vector, by each word's importance in the text corpus. In detail, TF-IDF is calculated on top of two variables: term frequency by Equation 1 and inverse document frequency by Equation 2. The former one computes the relative frequency of each term (or word) while the later one measures the information density of each term. Equation 3 is applied to compute the weight of each term. We set the minimum document frequency to 3 and the max features (or terms) to 10000. Below lists the equations:

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t'\in d} f_{t',d}} \tag{1}$$

$$\mathrm{idf}(t,D) = \frac{N}{|\{d \in D : t \in d\}|} \tag{2}$$

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d)\cdot\mathrm{idf}(t,D) \tag{3}$$

where $f_{t,d}$ is count of a term t in a document d, D is the set of documents, and N is the total number of documents ($N = |D|$).

After that, we applied k-means clustering on the numerical vectors by varying k from 2 to 8, and found the best result is achieved when $k = 4$. For the implementation details, our code is written in Python, using libraries including NLTK WordNetLemmatizer, sklearn TfidfVectorizer, and sklearn KMeans. Algorithm 1 lists the main steps of the clustering method.

Here we visualize the clustering result with PCA (Principal Component Analysis) under two components, which is shown in Figure 3. It turns out that though some clusters can be relatively well separated (e.g., cluster 2 and 3), the overlap is still prominent for some clusters (e.g., cluster 1 and 4). The silhouette score computed on the numerical vectors, which is a common metric to evaluate the quality of the clusters, is only 0.0110, indicating achieving good clustering results on the posts might be very hard. Hence, we decide to adopt an alternative approach: sampling representative posts from each cluster and applying keyword-based matching to tag each post. Though manual analysis is involved in this approach, we are able to obtain fine-grained control over the results (i.e., reducing false positives[1] and false negatives[2]), without examining all the posts.

Algorithm 1 The Pseudo-Code of the Clustering Method

```
Input: The raw document corpus $C$, the term corpus
       $C_{\mathrm{word}}$, the minimum document frequency $\mathrm{min}_{\mathrm{df}}$,
       the max feature number $\mathrm{max}_{\mathrm{f}}$, cluster number $k$
Output: the cluster labels $L$
$i = 0$;
if $i < \mathrm{len}(C)$ then
    Seq = segment($C[i]$.strip().lower());
    $C_{\mathrm{word}}[i] = []$;
    $j = 0$;
    if $i < \mathrm{len}(\mathrm{Seq})$ then
        if Seq[$j$] is noun, verb, or adjective then
            $C_{\mathrm{word}}[j]$.append(Seq[$j$]);
        $j = j + 1$;
    $i = i + 1$;
$V$ = TfidfVectorizer($\mathrm{min}_{\mathrm{df}}$, $\mathrm{max}_{\mathrm{f}}$).fit($C_{\mathrm{word}}$);
$L$ = Kmeans($k$).fit($V$);
return $L$;
```

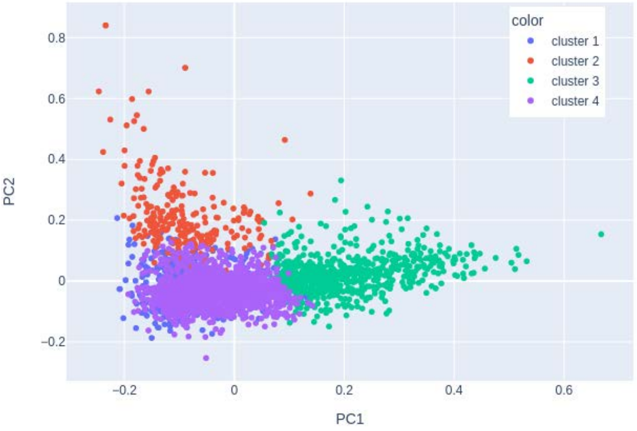


FIGURE 3. PCA visualization of the k-means clustering result of the Cloudflare posts. PC1 and PC2 are two main components. We set k to 4.

Though one can implement a simple keyword matching method, and tag a post when a uni-gram keyword or a bi-gram keywords pair is matched, we found this method led to the imprecise result, because a large number of high-frequency keywords have ambiguous meanings. Hence, we propose an adaptive keyword matching method. We manually decide the set of tags that are used to label a post, and associate a group of keywords with each tag. Then, we classify the keywords into two levels, with the first level to achieve good post coverage, and the second level for accurate tagging. When a post matches more than a threshold of level-1 keywords or one level-2 keyword, the post will be classified. We applied this approach to label the general categories of a post, and

[1] False positive means the wrong label is assigned.

[2] False negative means the ground-truth label is not assigned.

TABLE 1. The impact of *k* on k-means.

k	Silhouette Score
2	0.0084
3	0.0088
4	0.0110

TABLE 2. The comparison between DBScan and k-means.

Method	Silhouette Score
DBScan	-0.2514
Kmeans	0.0110

the result is described in Section IV (''General Categories of Posts'').

1) OTHER PARAMETERS AND METHODS

Here we describe the impact of different parameter values on the clustering result. We also discuss the different choices of methods.

We chose the k-means clustering method and set its parameter k to 4. In Table 1, we show the silhouette score under $k = 2, 3, 4$, and the score of $k = 4$ is 10x higher than the other k values. For the higher k values, though the silhouette score increases slightly (e.g., 0.0115 when $k = 5$), we found the clusters are harder to discern under PCA visualization.

In addition to k-means, another popular clustering method is DBScan. In Table 2, we compare the silhouette score between DBScan and k-means. For DBScan, we set the maximum distance and minimum data points in a cluster to 0.6 and 2, by exploring different parameter values, but its silhouette score is still 0.26 lower.

We use TF-IDF to translate a word sequence into a numerical sequence. Another popular approach is word embedding, which learns the vectorized representation with neural networks. We tried this approach but the result is unsatisfactory. First, we found the volume of the crawled forum posts is insufficient to train an accurate neural network for this purpose. Second, we also tried to fine-tune a pre-trained NLP model from Google [47] with a subset of posts labeled by us, but the result is still unsatisfactory, due to that the subset is small and it is time-consuming to label posts. On the contrary, TF-IDF does not rely on a large training dataset.

IV. MEASUREMENT RESULTS

A. STATISTICS OF THE CRAWLED FORUMS

Table 3 shows the number of posts, replies, and the dates of the first and last crawled posts of each forum. To notice, for Cloudflare, there is a pre-processing step to filter out the non-DNS posts (described in Section III, ''Forum Selection''), and Table 3 shows the numbers before and after this step. All the following measurement tasks performed on Cloudflare are done on the filtered data. We downloaded about 2 years' data from Cloudflare and Spiceworks, and 2 and a half years' data from OpenDNS, which achieve good coverage of their active and archived posts. For Comodo, we downloaded the posts from 2010 to 2019 (its last post was created in Nov. 2019), because the volume of posts per year is small (only dozens). Overall, Cloudflare is most active, with 11705 posts and 47474 replies observed in the study period, which are more than 10x of any other forum and can be attributed to its large user base. We found the engagement from users are active for all target forums, with the reply-to-post ratio (i.e., #Replies/#Posts from Table 3) spanning from 3.2 (OpenDNS) to 6.8 (Comodo).

TABLE 3. Statistics about the crawled DNS forums. ''Cloudflare-UF'' means unfiltered Cloudflare data.

Forum	#Posts	#Replies	Start	End
Cloudflare-UF [42]	32151	145123	Jan. 2020	Dec. 2021
Cloudflare [42]	11705	47474	Jan. 2020	Dec. 2021
OpenDNS [49]	566	1806	Apr. 2019	Dec. 2021
Comodo [43]	272	1838	Jul. 2010	Nov. 2019
Spiceworks [44]	790	5253	Jan. 2020	Dec. 2021

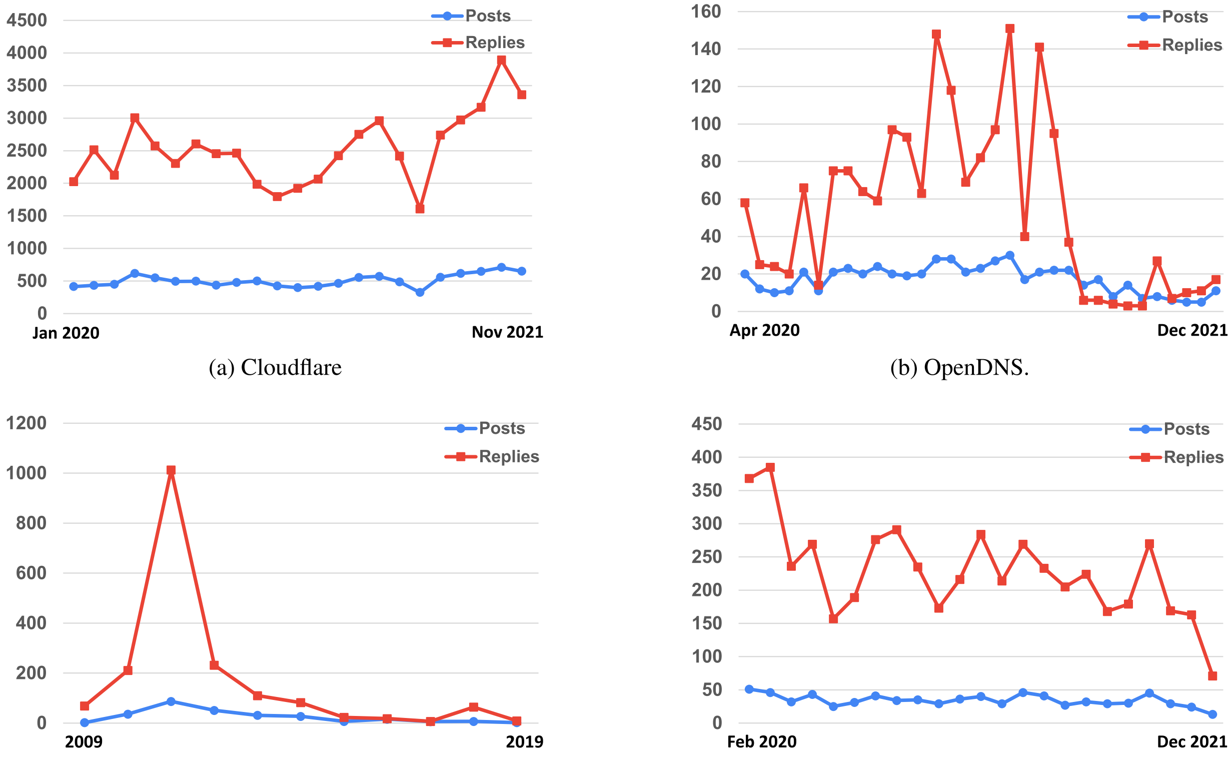
In Figure 4, we also show the number of posts and replies of each forum by years (Comodo) or months (Cloudflare, OpenDNS, and Spiceworks). Overall, the number of posts is stable across different months or years (e.g., ranging from 400 to 700 per month for Cloudflare), showing DNS users have encountered operational issues constantly. On the other hand, the number of replies has a much wider range (e.g., ranging from 1,600 to 3,800 per month for Cloudflare), suggesting some issues have triggered intensive discussion from DNS users. In Section V, we present some examples of such posts.

In Figure 5, we count the frequencies of words showing in the forum posts and visualize the top words with TagCrowd [48], in order to highlight the major concerns from users in DNS. It turns out some issues are persistent among different service providers, related to keywords like ''ip'', ''http'', ''server'', ''domain'', and ''website'', showing a large number of DNS issues are related to domain resolution, domain management, and web visits. On the other hand, we also found some top keywords that are unique to a DNS forum, e.g., ''ssl'' for Cloudflare, ''filter'' and ''block'' for OpenDNS, ''secure'' for Comodo. We speculate this is because these DNS providers also support other DNS-related services, e.g., CA for Cloudflare, and domain filter by OpenDNS and Comodo.

B. STATISTICS ABOUT THE FORUM USERS

Next, we count the posts and replies made by users to measure the activities of the users of each forum. Table 4 shows the number of users who have made posts and replies, and the average number of posts and replies per user. It turns out on average a user only wrote about 1 post (i.e., 1.1 for Cloudflare and OpenDNS, 1.2 for Comodo and Spiceworks), but the ratios between replies and users are much higher (i.e., 5.9, 4.7, 3.1 and 4.4 for Cloudflare, OpenDNS, Comodo, and Spiceworks).

Then we take a closer look at users' activities, and list the top 5 users with the most posts and replies



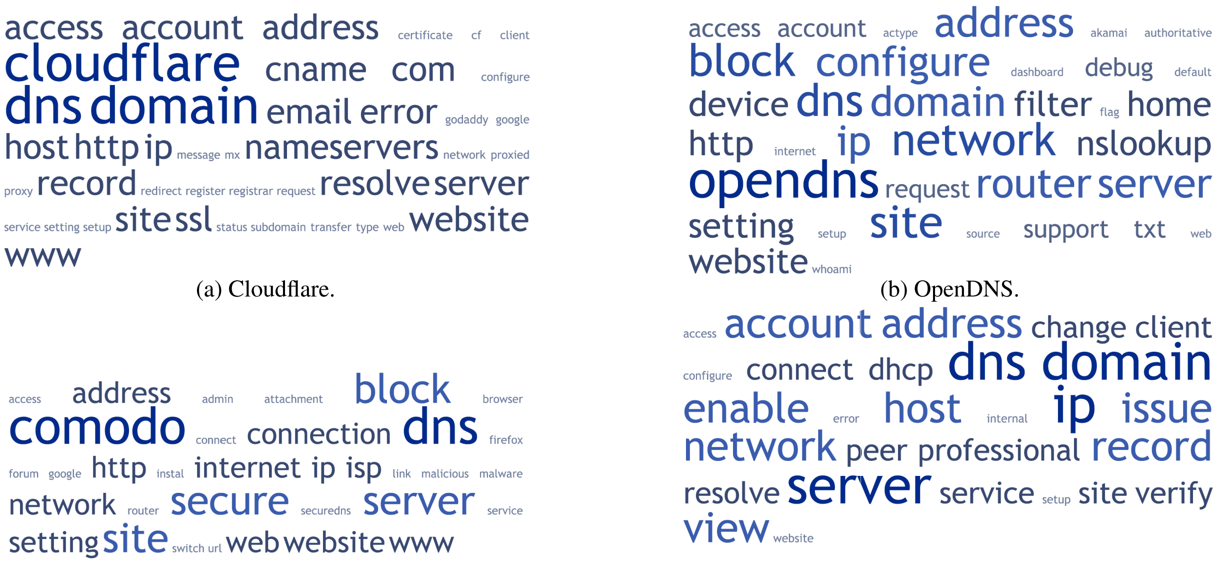
(a) Cloudflare

(b) OpenDNS.

(c) Comodo.

(d) Spiceworks.

FIGURE 4. The number of posts and replies per month (for Cloudflare, OpenDNS, and Spiceworks) or year (for Comodo). The last month of Cloudflare is not shown because we stop crawling Cloudflare in the middle of December 2021.


access account address certificate cf client
cloudflare cname com configure
dns domain email error godaddy google
host http ip message mx nameservers network proxied
proxy record redirect register registrar request resolve server
service setting setup site ssl status subdomain transfer type web website
www
access account actype address akamai authoritative
block configure dashboard debug default
device dns domain filter flag home
http internet ip network nslookup
opendns request router server
setting setup site source support txt web
website whoami
access address admin attachment block browser
comodo connect connection dns firefox
forum google http instal internet ip isp link malicious malware
network router secure securedns server service
setting site switch url web website www
access account address change client
configure connect dhcp dns domain
enable error host internal ip issue
network peer professional record
resolve server service setup site verify
view website

(a) Cloudflare.

(b) OpenDNS.

(c) Comodo.

(d) Spiceworks.

FIGURE 5. Tag clouds generated from the forum posts. The tags with larger fonts have higher frequency.

in Table 5 and Table 6. For the users who wrote the posts (i.e., post authors), though Cloudflare has seen some quite active users (e.g., No.1 and No.2 users made 46 and 21 posts), the top users are much less active in other forums (e.g., the No.1 user of Comodo and Spiceworks only made 8 and 7 posts). On the other hand, every forum has a few users who are very active and resolve a lot of post authors' problems. For example, the top 1 and 2 users in Cloudflare have written 8026 and 5252 replies, accounting for *16.9%* and *11.1%* of all Cloudflare replies. The top 1 user in OpenDNS even

TABLE 4. The number of users who made posts ("Posted"), replied ("Replied"), and the average posts and replies per user ("Avg. Posted" and "Avg. Replied").

#Users	Cloudflare	OpenDNS	Comodo	Spiceworks
Posted	10211	514	224	653
Avg. Posted	1.1	1.1	1.2	1.2
Replied	8050	388	592	1193
Avg. Replied	5.9	4.7	3.1	4.4

TABLE 5. The top 5 users ranked by the number of posts they wrote.

Ranking	Cloudflare	OpenDNS	Comodo	Spiceworks
1	46	17	8	7
2	21	7	5	6
3	16	4	4	5
4	11	3	3	5
5	11	3	3	4

TABLE 6. The top 5 users ranked by the number of replies they made.

Ranking	Cloudflare	OpenDNS	Comodo	Spiceworks
1	8026	930	88	221
2	5252	41	88	140
3	1554	24	75	127
4	1028	14	55	104
5	897	13	50	95

TABLE 7. The number of days and replies (max, min, average and median) between the creation of a post till it is solved.

	Cloudflare		Spiceworks	
	#Days	#Replies	#Days	#Replies
Max	386	165	62	47
Min	0	1	0	1
Average	0.63	4.51	1.58	5.91
Median	0	3	0	4

contributed to *more than 50%* of all replies (930 over 1806). We suspect some of those users are actually service maintainers or developers, and they play critical roles in addressing DNS operational issues.

C. ISSUE RESOLUTION

We found Cloudflare and Spiceworks allow the post author or the forum admin to mark a reply as the solution. Hence, with the timestamps on each post and reply, we can assess the duration between when an issue was raised and when it was resolved. Cloudflare and Spiceworks mark a reply as "Solution" and "Best Answer" with an extra HTML field, hence, we studied Cloudflare and Spiceworks for this measurement task. Table 7 shows the statistics of the resolved posts. In particular, for a post with a solution reply, we count the number of days and replies (including the solution reply) and compute the aggregated statistics.

For Cloudflare, we found 4446 posts were marked as solved, which only account for 37.9% of all the 11705 posts. The average days and replies are 0.63 and 4.51, while the median days and replies are 0 and 3, suggesting most of the raised issues can be resolved quickly (in a day, with a few replies). On the other hand, we found in the extreme case, resolving an issue could take more than a year, with more than 100 replies. For Spiceworks, we found 239 posts were marked as solved, accounting for 30.2% of all the 790 posts. Similar as Cloudflare, resolving an issue often takes a short period (e.g., in average 1.58 days and 5.91 replies, and the median values are 0 and 4).

D. GENERAL CATEGORIES OF POSTS

Based on our empirical analysis of the posts sampled from each cluster (see Section III-D), we found the issues are mainly about failures in domain management (e.g., errors at nameserver configurations and domain registrations), network connectivity (e.g., the DNS requests from clients timeout), security (e.g., DNS communication is hijacked and domain filtering is applied for clients' security), and website (e.g., websites cannot be visited because their domains are not resolving). We applied the keyword matching algorithm described in Section III-D (the keywords are listed in Table 8) and labeled each post by the 4 general categories, and the majority of posts can be tagged (the untagged posts for Cloudflare, OpenDNS, Comodo and Spiceworks are only 11%, 25%, 10% and 13%). The results are shown in Table 9. Noticeably, our categories are not exclusive, so a post can have multiple tags. The tags generated by the follow-up tasks are also non-exclusive.

To validate the accuracy of our adaptive keyword matching method, we sampled 150 posts from Cloudflare (100 posts have at least one category, and 50 posts have no category), and manually examined their categories. We found 4 posts were assigned with the wrong category and 1 post did not get a category (but it should), which shows our method can achieve good accuracy.

We found for each forum, all general categories have a significant amount of posts: the lowest are Security for Cloudflare (24.8% posts), Security for OpenDNS (22.1% posts), Domain Management for Comodo (14% posts), and Security for Spiceworks (28.5% posts). Yet, the top concern for each forum actually differs, with Website for Cloudflare (58.8%), Website for OpenDNS (47%), Security for Comodo (62.1%), and Network Connectivity for Spiceworks (60.8%).

E. RECORD TYPE

In addition to classifying the posts by the general categories, we also tag the posts by DNS-related features, like the record type. In particular, we match each post using keywords of the known record types [50] (e.g., "A record" and "CNAME record"). The number of posts by record type is shown in Table 10.

It turns out a non-negligible ratio of posts (e.g., 3703 out of 11705 Cloudflare posts) mentioned record types, and the common record types, including A, AAAA, CNAME, NS, MX, SOA, SRV, TXT, and PTR are also mentioned most. Though A record should appear most in a zone file, other records, like CNAME for Cloudflare (mentioned in 1462 posts) and TXT for OpenDNS (mentioned in 54 posts) have caused many issues. For most of the records, one

TABLE 8. The keywords used to label a post's general category.

Category	Level	Keywords
Network Connectivity	1	"slow", "slowly", "fail", "failure", "block", "time", "traffic", "route", "packet", "refuse", "masquerade", "transmit", "performance", "load", "redirect", "stable", "limit", "handshake", "404", "tcp", "udp", "pagespeed" "connect", "resolve", "access", "accessible", "tunnel", "forward", "network", "socket", "websocket", "websockets"
	2	"connectivity", "connection", "timeout", "resolver", "resolvers", "time out", "rate limit", "packet loss", "refuse connection", "connection failure", "connectivity failure", "redirect loop"
Domain Management	1	"account", "switch", "migrate", "expire", "lookup", "dnssec", "renew", "renewal", "update", "portfolio", "manage", "management", "iam", "configure", "configuration", "subdomain", "subdirectory", "setting", "setup", "glue", "misconfigure", "email", "mail", "record", "propagate", "propagation", "apex"
	2	"registration", "registrar", "register", "registry", "nameserver", "nameservers", "transfer", "mydomain", "whois", "godaddy", "icann", "iana", "bluehost", "go daddy", "name server", "primary domain", "main domain"
Security	1	"block", "filter", "expose", "forbid", "captcha", "safety", "safe", "unsafe", "insurance", "guarantee", "warrant", "ward", "shelter", "https", "proxy", "proxied", "feign", "intercept", "cert", "certificate", "bot", "robot", "dnssec", "waf", "provision", "encrypt", "unencrypt", "verify", "unverified", "intercept", "mitigation", "mechanism", "onion", "protect"
	2	"security", "secure", "insecurity", "insecure", "firewall", "ddos", "dos", "hijack", "ssl", "tls", "malware", "attack", "cyberattack", "cybersecurity", "authentication", "authenticate", "phishing", "credential stuffing"
Website	1	"display", "redirect", "bypass", "browser", "load", "response", "responsiveness", "visual", "view", "visitor" "optimize", "optimization", "setup", "cdn", "cert", "certificate", "bot", "robot", "crawler", "admin", "http", "https", "page", "brick", "deploy", "cpanel"
	2	"site", "website", "webpage", "wordpress"

TABLE 9. The number of post under each general category. "Conn." and "Sec." mean "Connectivity" and "Security".

Forum	Management	Conn.	Sec.	Web	Unknown
Cloudflare	5893	4315	2908	6877	1319
OpenDNS	162	244	125	266	142
Comodo	38	113	169	153	27
Spiceworks	441	480	225	376	102

TABLE 10. The number of posts by record types.

Record	Cloudflare	OpenDNS	Comodo	Spiceworks
A	1118	0	2	118
AAAA	217	1	0	12
CNAME	1462	1	0	44
CAA	22	0	0	0
CERT	169	2	0	16
DNSKEY	16	0	0	0
LOC	22	0	0	0
NS	647	1	1	20
MX	522	0	0	45
SOA	52	0	1	8
SRV	131	0	0	23
TXT	448	54	0	24
PTR	47	1	0	34
AFSDB	1	0	0	0
CDNSKEY	2	0	0	0
DNAME	2	0	0	0
RP	2	0	0	0
RRSIG	4	0	0	0
SSHFP	2	0	0	0
NSEC	1	0	0	0

major issue that a user encountered is that the record is not updating or propagating to other DNS servers/resolvers, and the root causes include conflicting with other records, wrong format, etc.

It is also interesting to see some uncommon record types have been mentioned at a non-negligible frequency (i.e., more than 10 posts), e.g. CAA record (specifies CAs that are authorized to issue certificates for the domain), CERT record (stores certificates and related revocation lists), LOC record (stores geographical location information), and DNSKEY (hold a public key that verifies a DNSSEC record). For CAA, CERT, and DNSKEY issues, the main root cause is that the domain owner who wants to support encryption fails to configure their DNS correctly.

F. STATUS CODE

We found for certain service providers, error codes are attached to the responses to help DNS users and domain owners for debugging. One example is Cloudflare, which uses error code 1xxx to help the owners and visitors of the websites proxied by Cloudflare for troubleshooting [51]. Here we measure the frequencies of error codes mentioned in the Cloudflare posts to infer the major issues encountered by DNS users and domain owners, which are shown in Table 11.

In total, we found 83 posts include error codes, and 57 (68.6%) of them had the error code 1004, which suggests either the domain violated the terms of service of the Cloudflare proxy or DNS changes have not yet propagated. Our analysis of record types also shows the major issue is that DNS changes are not updating. Error 1001 ranked second, and its usual root cause is the misconfiguration of CNAME by the domain owner (e.g., the target of CNAME does not resolve), which accords with our finding in Table 10 that CNAME has the most mentions (1462 posts). On the other hand, it is surprising that less than 100 posts mentioned error codes, which should provide important hints for issue diagnosis.

G. LOCATIONS MENTIONED IN ISSUES

As illustrated in Figure 1, DNS resolution can be impacted by a number of network locations (e.g., stub resolver, forwarder,

TABLE 11. The number of Cloudflare posts by error codes.

Error	#Posts	Explanation
1001	9	DNS resolution error
1002	1	DNS points to prohibited IP
1003	1	Access Denied: Direct IP Access Not Allowed
1004	57	Host Not Configured to Serve Web Traffic
1010	4	Website owner banned browser access by signature
1020	5	Access denied
1034	1	Edge IP Restricted
1037	1	Invalid rewrite rule (failed to evaluate expression)
1041	4	Invalid request rewrite (invalid header value)

TABLE 12. The number of posts mentioning network locations. "ADNS", "Recursive R." and "Stub R." mean authoritative nameserver, recursive resolver (both ISP and public), and stub resolver.

Location	Cloudflare	OpenDNS	Comodo	Spiceworks
Registrar	1754	1	2	46
Root	339	3	2	52
TLD	109	0	0	4
SLD	2	0	0	0
ADNS	35	0	0	4
Firewall	286	7	16	91
Gateway	1098	32	8	48
Router	206	161	27	62
Forwarder	383	11	1	143
Recursive R.	7	1	3	7
Stub R.	8	0	0	7

recursive resolver, authoritative nameservers, and registrars). We are interested in the distribution of locations in DNS operational issues and try to infer it by searching for the posts with a keyword list of network locations. Multiple search keywords are mapped to one location (e.g., client resolver and stub resolver) if they have similar meanings. Table 12 shows the number of posts mentioning a network location.

It turns out that for Cloudflare, most users mention the registrar when encountering issues (1754 posts), which can be explained by the fact that Cloudflare also runs domain registration service [52]. For the other forums, interestingly, we found the network components that are on the path between DNS clients and DNS servers, including router, gateway, and firewall, have raised more issues than other core DNS components (e.g., recursive resolver). For instance, 161 and 27 posts in OpenDNS and Comodo mention router. DNS forwarder is designed to simply pass the DNS request to the next-hop without resolving, but still, it encounters a lot of issues (e.g., 143 posts in Spiceworks mention forwarder).

V. EXEMPLAR POSTS

In this section, we perform qualitative analysis on individual posts to understand the particular problems users have encountered and how they are resolved. We first select the posts that have engaged many replies, suggesting either the related issues were encountered by many users or the problem resolution is uneasy. Then, we select the posts about the issues in IPv6, through which we want to highlight users' experiences in adopting IPv6.

A. POSTS WITH MANY REPLIES

For each forum, we sort the posts by the total number of replies and pick the top 5 posts. We summarize two posts per forum, and their titles and number of replies are shown in each box below.

Cloudflare (#Replies:121): Massive SSL/TLS mess - Cloudflare error page still after removing my site from Cloudflare service completely.

The user who owned a website hosted by Bluehost tried to switch to Cloudflare to remove the support of the deprecated TLS 1.0 and 1.1 (Cloudflare forced TLS 1.2 at a minimum while Bluehost refused to remove support for TLS 1.0 and 1.1). The user changed A record to point to Cloudflare and also installed new certificates on his/her server. However, the user kept seeing Cloudflare error code 403 (forbidden) and 526 (invalid SSL certificate), no matter how the user switched between Cloudflare and Bluehost and changed certificate settings. The issue was partially resolved when the user realized he/she set a wrong IP address in the A record, which does not belong to Cloudflare. Later, the user identified Bluehost has an integration issue with Cloudflare, which keeps using the wrong certificate. The feature is called "auto-SSL", and by turning off this feature, the issue was finally resolved.

Here we summarize a few interesting observations after reading through the replies. 1) Initially, a user (a Cloudflare MVP) asked about the domain name to diagnose the issue, but the post author refused to list the domain name for privacy concerns. Later, the user had to post a reply with the domain name and delete it quickly (the Cloudflare MVP was informed ahead). After that, the issue of 403 was quickly resolved when the Cloudflare MVP learnt the website IP address. 2) A number of replies from the post author complained the webpage is not updated after making the changes, which is actually caused by the long TTL set by DNS resolvers and cache. 3) The post author also mentioned Bluehost technicians kept blaming Cloudflare but the issue seems to be caused by the Bluehost's feature ("auto-SSL").

Cloudflare (#Replies:72): Unable to update DDNS using API for some TLDs.

In April 2020, the post author reported that Cloudflare disallowed DDNS (Dynamic DNS, which updates DNS records automatically for IP address changes) users to update their domain settings with Cloudflare API, when the domain is hosted under free TLDs, like.tk. Later, it was identified that the user has to either use Cloudflare dashboard or upgrade their account. This post was replied by more than a dozen of users, who were upset about this policy change and the bad communication from Cloudflare. Some users speculate Cloudflare made this change to counter spammers who have been known to host malicious domains on the free TLDs like.tk [53], but the follow-up replies all argue they need domains under free TLDs for legitimate purposes.

OpenDNS (#Replies:49): Opendns updater 2.2.1 Error Message "Looks like there's no internet connectivity".

OpenDNS Updater is a DNS client software helping OpenDNS users update their IP addresses associated with their registered DDNS domains. The error message encountered by the post author shows the updater cannot connect to the Internet. The root cause inferred by an expert user is that the router of the post author cannot connect to the correct OpenDNS server IP. However, the debugging process is lengthy. The expert user asked for the screenshot with the OpenDNS address configuration, and asked the post author to run debugging commands (e.g., "nslookup -type=txt debug.opendns.com") and report the output, but it takes quite some replies for the post author to take the right screenshots and report the right output. A few other users replied they have the same issue and posted their screenshots, but no one helped them further.

OpenDNS (#Replies:21): Not blocking adult sites.

OpenDNS has a feature to block access to adult websites, when it is configured as the resolver for a user. However, the post author found it is not blocking access to some well-known adult websites. The root cause inferred by an expert user is that the user did not use the correct IPv6 addresses of OpenDNS to register his/her device on the OpenDNS dashboard, to enable this website blocking feature. Yet, it turns out this feature is not effective when the user enables it on his/her Android device and moves out of his/her home LAN. Similar to the previous post, the user is advised to run the same debugging commands and report the results for diagnosis.

Comodo (#Replies:601): Report Blocked Sites You Believe Are Safe Here

This post was created in 2011. Back then, Comodo SecureDNS used a blacklist to block users' access to malicious websites, by returning sink-holed IP addresses. To correct false positives, the users can post the blocked URLs and their scanning results (e.g., by URLVoid), and wait for the websites to be removed. It turns out a lot of websites were found benign in the end. Also, a few users complained that the website is still blocked after reporting, but it is often because their DNS cache is not cleared.

Comodo (#Replies:33): Stop using SecureDNS

The post author who installed Comodo SecureDNS asked for advice on changing DNS settings back, because SecureDNS blocked too many websites. The replies also suggest using commands (e.g., "ipconfig /flushdns") to clear DNS cache after changing the DNS configurations. Interestingly, the discussion started to detour in the middle to compare ISP resolvers and public resolvers, and a few users mentioned they are hesitant to use public resolvers like Comodo SecureDNS because they are slower than the ISP resolver (the post was made in 2011), even though security features might be provided.

Spiceworks (#Replies:61): Do you use non ISP DNS ?

The post author wanted to switch from an ISP resolver (e.g., Comcast) to a public resolver that can block malicious websites and malware, and asked for recommendations. In the replies, Google DNS, OpenDNS, and Quad9 were mostly mentioned. Less famous DNS software or service providers, like Pi-hole and Cleanbrowsing were also mentioned because they are able to filter advertisements. The replies in general believe public resolvers are better choices, because they support more features (e.g., one mentioned ISP resolvers do not support DNS-over-HTTPS). As this post was made in 2021, it indicates an increasing trend of adopting public resolvers, which is also echoed by other studies [54].

Spiceworks (#Replies:99): Dcdiag results have missing SRV records

The post author managed domain controllers (DC) in an enterprise. He/she saw an old Windows 2008 server failed, and removed it from the active directory (AD). However, the ERP system of the enterprise started to respond slowly. When the user executed dcdiag command, which is a Microsoft command-line tool for DC troubleshooting, errors like missing SRV records are reported. The replies suggested many solutions, like changing the forwarder settings, checking if the IP addresses are private, flushing DNS cache, and re-register DNS records on DC, but the problem was not solved till the post was locked. Noticeably, the post author reported the output of running the debugging tools like ipconfig, but private information is also listed (e.g., the hostanme, IP and MAC addresses of the enterprise machines).

B. POSTS ABOUT IPv6

When sampling the posts empirically, we found IPv6 has been mentioned at a notable frequency (e.g., 166, 28, 8, and 20 posts mentioned IPv6 in Cloudflare, OpenDNS, Comodo, and Spiceworks). Most of these posts turn out to be about the compatibility issues with IPv6 and some users are also confused about the IPv6 setup. Our analysis suggests DNS users' experiences with IPv6 should be improved, and we describe an exemplar post below.

OpenDNS (#Replies:7): Phones Apps Are Not Getting Filtered with IPv6 setup on eero WiFi

The post author found OpenDNS does not block visits from mobile apps to certain domains (e.g., YouTube) when IPv6 is enabled by the local WiFi, but the block is still effective on the desktop devices. The replies show OpenDNS does not provide content filtering when IPv6 address is used, and it is even suggested by the other users to turn off IPv6.

VI. DISCUSSION

A. SUMMARY OF THE FINDINGS

We summarize the major findings from Section IV and Section V. 1) We observed different levels of activities among DNS forums (e.g., more than 10000 posts from Cloudflare but only 272 posts from Comodo), but a post often got multiple replies (ranging from 3.1 to 5.9 on average) across forums. 2) Though not all posts were marked as resolved (37.9% for Cloudflare and 30.2% for Spiceworks), resolving an issue usually takes a short period of time (on average less than 2 days). 3) All the forums have highly active expert users to answer troubled users (e.g., the top replier in Cloudflare makes more than 8000 replies). 4 The general concerns were centered around network connectivity, domain management, security, and websites across DNS forums. 5 DNS failures were were reportedly associated with 20 different record types, and configuration issues in the DNS zone have become a major root cause 1462 posts in Cloudflare were about CNAME). 6 In terms of network locations, the core DNS components (e.g., recursive resolver and stub resolver have reported much fewer issues than other components like registrar (1754 for Cloudflare) and router (161 for OpenDNS). 7) With a qualitative study on exemplar posts, we found prominent concerns were actually raised about the auxiliary functionalities offered by DNS service providers (e.g., DDNS and domain blocking). While it is a common practice to instruct a user to run command-line tools and post the output, we found it could introduce privacy risks, e.g., private IP addresses and hostnames being posted. Also, the delay between applying a fix and issue-resolving can be long, due to the staled DNS cache, which tends to confuse users and domain confuse users and domain owners.

B. FINDINGS COMPARED TO PRIOR WORKS

Since prior works studied DNS failures from DNS logs [1], [4], a data source different from DNS forums posts, we compare our findings with theirs. 1) Regarding record type, Gao et a. [4] identified that A, AAAA, PTR and DNSBL records were the major sources of DNS failures in 2013, while the top-4 record types from our studied forums are CNAME, A, NS and MX (see Table 10). The differences can be attributed to that domain owners were covered by our study, so domain configuration errors were recorded, while prior works were solely about DNS end-users. Yang et al. [1] focused on the failures of A and AAAA, and found the error rate of AAAA record can be 9x from A record ($\frac{1-0.931}{1-0.358}$ from Table 1 of [1]). Our result shows AAAA is more error-prone as well, by considering the traffic volume ratio between A and AAAA: the ratio of posts between A and AAAA is 5.38x (adding all 4 forums together) in Table 10, while the traffic volume of A over AAAA is 8.29 ($\frac{0.862}{0.104}$ from Table 1 of [1]). 2) Regarding the locations of failures, Gao et al. showed that local resolvers, which represent client-side DNS infrastructure [55], have higher success rate compared to the root servers (66.9% compared to 46.0% in Table 3 of [4]. Our result in Table 12 shows the opposite: 396 posts are about root servers, but 2613 posts are about the client-side DNS infrastructure, by counting firewall, gateway, router, forwarder, recursive resolver and stub resolver all together. The high number of posts of the latter can be explained by that the client-side DNS infrastructure becomes more complex and error-prone in recent years or the local resolvers studied in [4] only covered part of client-side DNS infrastructure. 3) Both Gao et al. and Yang et al. pointed out a large number errors are relevant to TLDs (e.g., invalid TLD [4] and new gTLD [1]). Similarly, we found TLD has more posts than SLD and ADNS combined (113 compared to 41 in Table 12). 4) We acknowledge that forum-based analysis misses certain statistics like DNS TTL [4], as covered by DNS logs, but it also derives new insights like the errors related to domain management.

TABLE 13. Statistics about the crawled GitHub/GitLab issues.

Forum	#Posts	#Replies	Start	End
PowerDNS [57]	646	1963	Jan. 2020	Dec. 2021
Bind9 [58]	843	11131	Jan. 2020	Dec. 2021

C. RECOMMENDATIONS

We give a few recommendations that could improve the process of reducing and resolving DNS operational failures based on our observations. 1) We found a lot of troubled users' questions can be solved in a few replies, and many questions share the same root cause, which has been answered before. So far, these forums mostly rely on expert users to reply to each post, and we think this procedure can be partially automated by suggesting likely answers from the other posts to the troubled user (e.g., by computing the similarity between posts). 2) Directly pasting the output of the command-line tools might not be the best approach for failure debugging, due to the leakage of private information. The forums could develop input/output sanitizers to automatically mask the sensitive fields, e.g., the user's private source IP address, which should not prohibit the debugging process. 3) Since DNS cache update is passive (i.e., waiting for TTL to expire), users and domain owners were often confused about why the problems still have not been fixed even after they applied the right fix, which introduced a lot of unnecessary back-and-forth replies. We believe telling users to flush the cache (e.g., run ipconfig /flushdns) might not be sufficient, and active cache management can be considered, which can signal the cache updates more timely when the fixes are applied. Some proposals have been made about active cache updates to address performance bottleneck [56], they could be adapted to speed up failure recovery potentially.

D. LIMITATIONS OF THIS WORK

First, we focus on the forums of DNS service providers, but users who encountered operational issues could also ask the DNS software developers for help. We found a few highly popular DNS software share their code on GitHub (e.g., PowerDNS) and GitLab (e.g., Bind9), and the Issues under GitHub and GitLab could have relevant posts and replies. In fact, we have developed extra crawlers and mined the Issues under PowerDNS and Bind9, and the statistics are shown in Table 13. Yet, we did not perform in-depth analysis of these datasets because they have different metadata and usage compared to the studied forums. For example, a lot of the posts are about software bugs. We plan to investigate these datasets in the future and we will also release the crawled data in our repository.

Second, we chose 4 DNS forums as the study target, but admittedly there are many other DNS forums being left out. Table 14 lists some of the unstudied DNS forums of which we have found their URLs. We did not study them because

TABLE 14. Some of the DNS forums that were not crawled by us and the reasons.

Forum	Reason for No-crawl
Google [59]	Anti-crawling techniques deployed
AdGuard [60]	Less than 100 posts
Tencent DNS [61]	Mainly in Chinese
Ali DNS [62]	Mainly in Chinese
Yandex [63]	Less than 100 posts

they are difficult to crawl, have a small number of posts, are in non-English language, etc.

Third, we clustered the posts with unsupervised learning and created the keywords lists from the sampled posts. Errors in keyword matching are unavoidable, but through our sanity check, the accuracy of our approach is reasonably well, and we believe the conclusions drawn from the data are representative.

VII. CONCLUSION

In this paper, we conducted a systematic study of DNS operational issues. Different from related works that analyze DNS logs collected at resolvers and servers, we leverage the public posts presented at DNS forums to understand how DNS operational issues were caused and resolved. Particularly, we crawled 4 representative DNS forums and analyzed over 10000 posts and 50000 replies. We found DNS operational issues can be attributed to a broad range of factors and even components outside of core DNS functionalities. Privacy issues were also identified from the DNS posts. Overall, our study shows DNS forum did provide unique insights into DNS operational issues and we hope this study could help the DNS community in building more reliable DNS infrastructure.

REFERENCES

[1] D. Yang, Z. Li, and G. Tyson, "A deep dive into DNS query failures," in *Proc. USENIX Annu. Tech. Conf.*, 2020, pp. 507–514.

[2] J. Alois. (2020). *Going Down? Robinhood Experiences Another Period of Platform Outages*. [Online]. Available: https://www.crowdfundinsider.com/2020/06/162936-going-down-robinhood-experiences-another-period-of-platform-outages/

[3] R. Crozier. (2020). *Parliament IT Outage Caused by DNS Failure*. [Online]. Available: https://www.itnews.com.au/news/parliament-it-outage-caused-by-dns-failure-514302

[4] H. Gao, V. Yegneswaran, Y. Chen, P. Porras, S. Ghosh, J. Jiang, and H. Duan, "An empirical reexamination of global DNS behavior," in *Proc. ACM SIGCOMM Conf. SIGCOMM*, Aug. 2013, pp. 267–278.

[5] A. Barua, S. W. Thomas, and A. E. Hassan, "What are developers talking about? An analysis of topics and trends in stack overflow," *Empirical Softw. Eng.*, vol. 19, no. 3, pp. 619–654, 2014.

[6] C. Rosen and E. Shihab, "What are mobile developers asking about? A large scale study using stack overflow," *Empirical Softw. Eng.*, vol. 21, pp. 1192–1223, Jun. 2016.

[7] F. Calefato, F. Lanubile, M. C. Marasciulo, and N. Novielli, "Mining successful answers in stack overflow," in *Proc. IEEE/ACM 12th Work. Conf. Mining Softw. Repositories*, May 2015, pp. 430–433.

[8] K. Bajaj, K. Pattabiraman, and A. Mesbah, "Mining questions asked by web developers," in *Proc. 11th Work. Conf. Mining Softw. Repositories*, 2014, pp. 112–121.

[9] Y. Acar, M. Backes, S. Fahl, D. Kim, M. L. Mazurek, and C. Stransky, "You get where you're looking for: The impact of information sources on code security," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2016, pp. 289–305.

[10] F. Fischer, K. Bottinger, H. Xiao, C. Stransky, Y. Acar, M. Backes, and S. Fahl, "Stack overflow considered harmful? The impact of copy&paste on Android application security," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 121–136.

[11] UCI-DSP-Lab. (2022). *Project GitHub Repo*. [Online]. Available: https://github.com/uci-dsp-lab/dns_forum

[12] V. Pappas, Z. Xu, S. Lu, D. Massey, A. Terzis, and L. Zhang, "Impact of configuration errors on DNS robustness," in *Proc. Conf. Appl., Technol., Archit., Protocols Comput. Commun.*, 2004, pp. 319–330.

[13] A. Romao, "Tools for DNS debugging," FCCN, Universidade NOVA de Lisboa, Lisbon, Portugal, Tech. Rep., RFC 1713, Nov. 1994.

[14] Microsoft. (2018). *Description of the DNSLint Utility*. [Online]. Available: https://support.microsoft.com/en-us/help/321045/description-of-the-dnslint-utility

[15] Verisign. (2020). *DNSSEC Analyzer*. [Online]. Available: https://dnssec-analyzer.verisignlabs.com/

[16] V. Pappas, P. Fältström, D. Massey, and L. Zhang, "Distributed DNS troubleshooting," in *Proc. ACM SIGCOMM Workshop Netw. Troubleshooting, Res., Theory Oper. Pract. Meet Malfunctioning Reality*, 2004, pp. 265–270.

[17] S. K. R. Kakarla, R. Beckett, B. Arzani, T. Millstein, and G. Varghese, "GRooT: Proactive verification of DNS configurations," in *Proc. Annu. Conf. ACM Special Interest Group Data Commun. Appl., Technol., Archit., Protocols Comput. Commun.*, Jul. 2020, pp. 310–328.

[18] DNSChecker. (2020). *DNS Check Propagation Tool*. [Online]. Available: https://dnschecker.org/

[19] Check-Host. (2020). *Multi-Country Domain Resolving With DNS Service*. [Online]. Available: http://check-host.net/check-dns

[20] C. Lu, B. Liu, Z. Li, S. Hao, H. Duan, M. Zhang, C. Leng, Y. Liu, Z. Zhang, and J. Wu, "An end-to-end, large-scale measurement of DNS-over-encryption: How far have we come?" in *Proc. Internet Meas. Conf.*, Oct. 2019, pp. 22–35.

[21] N. Kephart. (2020). *Best Practices for Monitoring DNS*. [Online]. Available: https://www.thousandeyes.com/resources/dns-webinar

[22] J. Pang, J. Hendricks, A. Akella, R. De Prisco, B. Maggs, and S. Seshan, "Availability, usage, and deployment characteristics of the domain name system," in *Proc. 4th ACM SIGCOMM Conf. Internet Meas.*, 2004, pp. 1–14.

[23] D. Liu, S. Hao, and H. Wang, "All your DNS records point to us: Understanding the security threats of dangling DNS records," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2016, pp. 1414–1425.

[24] D. Liu, Z. Li, K. Du, H. Wang, B. Liu, and H. Duan, "Don't let one rotten apple spoil the whole barrel: Towards automated detection of shadowed domains," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2017, pp. 537–552.

[25] R. Potharaju, N. Jain, and C. Nita-Rotaru, "Juggling the jigsaw: Towards automated problem inference from network trouble tickets," in *Proc. 10th USENIX Symp. Networked Syst. Design Implement.*, 2013, pp. 127–141.

[26] W. Zhou, W. Xue, R. Baral, Q. Wang, C. Zeng, T. Li, J. Xu, Z. Liu, L. Shwartz, and G. Y. Grabarnik, "STAR: A system for ticket analysis and resolution," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 2181–2190.

[27] S. Agarwal, V. Aggarwal, A. R. Akula, G. B. Dasgupta, and G. Sridhara, "Automatic problem extraction and analysis from unstructured text in it tickets," *IBM J. Res. Develop.*, vol. 61, no. 1, pp. 4–41, 2017.

[28] S. P. Paramesh, C. Ramya, and K. S. Shreedhara, "Classifying the unstructured IT service desk tickets using ensemble of classifiers," in *Proc. 3rd Int. Conf. Comput. Syst. Inf. Technol. Sustain. Solutions (CSITSS)*, Dec. 2018, pp. 221–227.

[29] J. Han, K. H. Goh, A. Sun, and M. Akbari, "Towards effective extraction and linking of software mentions from user-generated support tickets," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2018, pp. 2263–2271.

[30] A. Revina, K. Buza, and V. G. Meister, "IT ticket classification: The simpler, the better," *IEEE Access*, vol. 8, pp. 193380–193395, 2020.

[31] Q. Wang, W. Zhou, C. Zeng, T. Li, L. Shwartz, and G. Y. Grabarnik, "Constructing the knowledge base for cognitive IT service management," in *Proc. IEEE Int. Conf. Services Comput. (SCC)*, Jun. 2017, pp. 410–417.

[32] S. Silva, R. Pereira, and R. Ribeiro, "Machine learning in incident categorization automation," in *Proc. 13th Iberian Conf. Inf. Syst. Technol. (CISTI)*, Jun. 2018, pp. 1–6.

[33] Q. Wang, T. Li, S. Iyengar, L. Shwartz, and G. Y. Grabarnik, "Online it ticket automation recommendation using hierarchical multi-armed bandit algorithms," in *Proc. SIAM Int. Conf. Data Mining*, 2018, pp. 657–665.

[34] J. Xu and R. He, "Expert recommendation for trouble ticket routing," *Data Knowl. Eng.*, vol. 116, pp. 205–218, Jul. 2018.

[35] N. Zhao, P. Jin, L. Wang, X. Yang, R. Liu, W. Zhang, K. Sui, and D. Pei, "Automatically and adaptively identifying severe alerts for online service systems," in *Proc. IEEE Conf. Comput. Commun.*, Jul. 2020, pp. 2420–2429.

[36] J. Han and A. Sun, "DeepRouting: A deep neural network approach for ticket routing in expert network," in *Proc. IEEE Int. Conf. Services Comput. (SCC)*, Nov. 2020, pp. 386–393.

[37] J. Garcia, Y. Feng, J. Shen, S. Almanee, Y. Xia, and A. Q. A. Chen, "A comprehensive study of autonomous vehicle bugs," in *Proc. ACM/IEEE 42nd Int. Conf. Softw. Eng.*, Jun. 2020, pp. 385–396.

[38] M. J. Islam, G. Nguyen, R. Pan, and H. Rajan, "A comprehensive study on deep learning bug characteristics," in *Proc. 27th ACM Joint Meeting Eur. Softw. Eng. Conf. Symp. Found. Softw. Eng.*, Aug. 2019, pp. 510–520.

[39] K. Man, Z. Qian, Z. Wang, X. Zheng, Y. Huang, and H. Duan, "DNS cache poisoning attack reloaded: Revolutions with side channels," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2020, pp. 1337–1350.

[40] X. Zheng, C. Lu, J. Peng, Q. Yang, D. Zhou, B. Liu, K. Man, S. Hao, H. Duan, and Z. Qian, "Poison over troubled forwarders: A cache poisoning attack targeting DNS forwarding devices," in *Proc. 29th USENIX Secur. Symp.*, 2020, pp. 577–593.

[41] R. Radu and M. Hausding, "Consolidation in the DNS resolver market—How much, how fast, how dangerous?" *J. Cyber Policy*, vol. 5, no. 1, pp. 46–64, Jan. 2020.

[42] Cloudflare. (2022). *Cloudflare Community*. [Online]. Available: https://community.cloudflare.com/

[43] Comodo. (2022). *Help—DNS*. [Online]. Available: https://forums.comodo.com/help-dns-b238.0/

[44] Spiceworks. (2022). *DNS Networking Forum*. [Online]. Available: https://community.spiceworks.com/networking/dns

[45] N. Li and D. D. Wu, "Using text mining and sentiment analysis for online forums hotspot detection and forecast," *Decis. Support Syst.*, vol. 48, no. 2, pp. 354–368, 2010.

[46] K. Kanagavalli and S. Tharani, "Analysing user posts for web forum using K-means clustering," *Int. J. Sci. Res. Publications*, vol. 4, no. 5, pp. 1–4, 2014.

[47] Google. (2022). *Cloud Natural Language*. [Online]. Available: https://cloud.google.com/natural-language

[48] D. Steinbock. (2022). *TagCrowd: Create Your Own Word Cloud From Any Text*. [Online]. Available: https://tagcrowd.com/

[49] OpenDNS. (2022). *OpenDNS Community—Community Help*. [Online]. Available: https://support.opendns.com/hc/en-us/community/topics/201091007-OpenDNS-Community-Community-Help

[50] SimpleDNS. (2022). *DNS Record Types*. [Online]. Available: https://simpledns.plus/help/dns-record-types

[51] Cloudflare. (2022). *Troubleshooting Cloudflare 1XXX Errors*. [Online]. Available: https://support.cloudflare.com/hc/en-us/articles/360029779472-Troubleshooting-Cloudflare-1XXX-errors

[52] Cloudflare. (2022). *Cloudflare Registrar New Domain Registration*. [Online]. Available: https://www.cloudflare.com/products/registrar/

[53] Spamhaus. (2022). *The 10 Most Abused Top Level Domains*. [Online]. Available: https://www.spamhaus.org/statistics/tlds/

[54] T. V. Doan, J. Fries, and V. Bajpai, "Evaluating public DNS services in the wake of increasing centralization of DNS," in *Proc. IFIP Netw. Conf. (IFIP Netw.)*, Jun. 2021, pp. 1–9.

[55] K. Schomp, T. Callahan, M. Rabinovich, and M. Allman, "On measuring the client-side DNS infrastructure," in *Proc. Conf. Internet Meas. Conf.*, Oct. 2013, pp. 77–90.

[56] E. Cohen and H. Kaplan, "Proactive caching of DNS records: Addressing a performance bottleneck," *Comput. Netw.*, vol. 41, no. 6, pp. 707–726, Apr. 2003.

[57] PowerDNS. (2022). *Issues PowerDNS/pdns*. [Online]. Available: https://github.com/PowerDNS/pdns/issues

[58] Bind9. (2022). *Issues ISC Open Source Projects/BIND GitLab*. [Online]. Available: https://gitlab.isc.org/isc-projects/bind9/-/issues

[59] Google Public DNS. (2022). *Google Issue Tracker*. [Online]. Available: https://issuetracker.google.com/bookmark-groups/77752?pli=1

[60] AdGuard. (2022). *DNS AdGuard Forum*. [Online]. Available: https://forum.adguard.com/index.php?tags/dns/

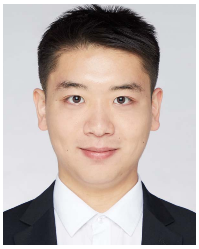
[61] Tencent. (2022). *Ask, Tencent Cloud*. [Online]. Available: https://cloud.tencent.com/developer/ask?q=hot

[62] AliDNS. (2022). *AliDNS Forum*. [Online]. Available: https://developer.aliyun.com/profile/gemj5geolfaqa

[63] XDA Forums. (2022). *Yandex DNS*. [Online]. Available: https://forum.xda-developers.com/tags/yandex-dns/

XIANRAN LIAO received the B.S. degree in computer science from the University of California at Irvine, Irvine, in 2022. His research interests include data mining and DNS security.

JIACEN XU (Student Member, IEEE) received the B.E. and master's degrees from Shanghai Jiao Tong University, China, in 2017 and 2020, respectively. He is currently pursuing the Ph.D. degree with the EECS Department, University of California at Irvine, Irvine. His research interests include data-driven and AI security.



QIFAN ZHANG (Student Member, IEEE) received the B.E. degree from ShanghaiTech University, China, in 2020. He is currently pursuing the Ph.D. degree with the EECS Department, University of California. His research interests include network security and machine learning security.

ZHOU LI (Senior Member, IEEE) received the Ph.D. degree in computer science from Indiana University Bloomington. He was a Principal Research Scientist at the RSA Laboratories, from 2014 to 2018. He is currently an Assistant Professor at the EECS Department, University of California at Irvine, Irvine. His research interests include cyber-security and privacy and machine learning. He received the NSF Career Award, in 2021.