Data in Brief 54 (2024) 110389

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

A user DNS fingerprint dataset

Josef Zápotocký [a], Jan Fiala [b], Jan Fesl [a,*]

[a] Department of Computer Systems, Faculty of Information Technology, Czech Technical University in Prague, Czech Republic
[b] Department of Applied Mathematics and Informatics, Faculty of Economics, University of South Bohemia in České Budějovice, Czech Republic

ARTICLE INFO

Article history:
Received 27 March 2024
Revised 2 April 2024
Accepted 3 April 2024
Available online 9 April 2024

Dataset link: A User DNS Fingerprint Dataset (Original data)

Keywords:
DNS
User
Machine learning
Identification
Fingerprint

ABSTRACT

Using a user DNS fingerprint allows one to identify a specific network user regardless of the knowledge of his IP address. This method is proper, for example, when examining the behavior of a monitored network user in more depth. In contrast to other studies, this work introduces a dataset for possible user identification based only on the knowledge of its DNS fingerprint created from the previously sent DNS queries.

We created a large dataset from the real network traffic of a metropolitan Internet service provider. The dataset was created from 2.3 billion DNS queries representing 6.2 million different domain names. The data collection took place over three months from 12/2023 to 02/2024.

The dataset contains a detailed user activity description in the sense of overall daily activity statistics and detailed 24 h activity statistics. Each dataset record contains a list of 1137 classification attributes. The absolutely unique feature of this data set is the classification of user activity based on categories of content accessed by a user.

The new dataset can be used for the creation of machine learning models, allowing the identification of a specific user

* Corresponding author.
E-mail address: fesljan@fit.cvut.cz (J. Fesl).

https://doi.org/10.1016/j.dib.2024.110389

without direct knowledge of their IP addresses or additional network location information. The dataset can also serve as a reference dataset for the creation of DNS fingerprints of users.

Specifications Table

Subject	Computer Networks and Communications Computer Science Machine Learning
Specific subject area	Encrypted data from a real network traffic collected within 3 months.
Type of data	Textual (CSV files).
How the data were acquired	The data was acquired as logs stored on a private DNS server. Further, CSV files containing the record of DNS user requests per day were assembled from the log records. Finally, an IP anonymization process was executed to guarantee the privacy of the real users - a unique hash value for each user to replace his IP address.
Data format	RAW
Description of data collection	The data was measured for approximately three months within the real computer network consisting of active network devices, i.e., routers and switches connected to the optic communication lines.
Data source location	Czech Technical University in Prague, Faculty of Information Technology Thákurova 9, Prague Czech Republic GPS: 50.105116930709194, 14.389857845702709
Data accessibility	Repository name: Zenodo Data identification number: 10.5281/zenodo.10887463 Direct URL to data: https://zenodo.org/records/10887463 The data set is freely available for usage without limitation.

1. Value of the Data

- The dataset [1] can be used to formulate a new vector form of a user's DNS identification fingerprint.
- The dataset contains sufficient samples to train machine learning models primarily designed to identify users based on DNS fingerprints.
- Data analysts or cybersecurity experts can use the dataset.
- The dataset can be a benchmark for comparing the quality of different algorithms or models designed to identify encrypted video streams.
- The use of the dataset lies in the possibility of speeding up the research, as it took three months and required one human to create it.
- The dataset is free of malicious content and is freely distributable.

2. Background

The main reason and motivation for creating this dataset was the possibility of training and validating different machine-learning models to identify users based on their DNS fingerprint because such a complex and thematically identical or similar freely downloadable dataset does not yet exist.

The current known methods of DNS fingerprint-based identification are machine learning, data mining and pattern mining, communication graph analysis, or statistical analysis, as mentioned in [2].

In the research work [3], semi-supervised learning is used with a dataset of DNS records. Unfortunately, it is not described in the paper whether it is DNS communication records or query records. They use algorithms such as the 1NN classifier to achieve the results of successful user re-identification with 55% accuracy and 87% re-identification accuracy.

In [4], the authors use a custom pattern miner to generate fingerprints from DNS records. They analyze streams of DNS query data where each stream comes from a single IP, build sets of domain names, and then continuously compare and evaluate these fingerprints based on their own definition of distances. Their user identification success rate is up to 91%.

The work of [5] focuses on the identification of infected machines. Here, the researchers present a feature list of observed attributes, consisting of DNS query records, from which they compute a DNS fingerprint. They use the MaxMind database to obtain domain information.

In [6], the authors focus on the domain names within the DNS queries sent, and based on their knowledge of these domain names, they determine the operating system of the clients with high accuracy.

The work [7] presents an innovative approach to fingerprinting using a custom dataset. They use a modified kNN classifier to create vectors from "epochs" and use vector matching with pattern mining. Their approach, implemented through machine learning, incorporates dataset statistics and is not limited to DNS queries.

In [8], the authors analyze periodic user behavior based on similarity between days, weeks, and months, achieving similarity between periods of up to 86% in the long term.

Recent work [9] specializes in generating DNS query descriptions for users. Researchers choose the ten most active IP addresses and perform experiments over them, tracking DNS lookups by a given IP address.

None of the works [2–9] that used DNS record datasets contained references to the datasets used, and this was one of the reasons for creating a dataset of their own.

3. Data Description

The data set was created from the DNS record queries stored in a log file of a DNS server located at a real metropolitan internet network service provider. The dataset contains a total of 2.3 billion DNS query responses belonging to 6.2 million unique domain names. The DNS queries were generated by more than four thousand unique users distinguished by IP addresses.

The period (12/2023-02/2024) was chosen because it includes many instances of individual behavior such as Christmas shopping, holidays and celebrations with multiple individuals, and normal user behavior during the work week and weekends. Fig. 1 depicts the pattern of a typical weekly flow of DNS queries sorted according to the day of the week and the hour of the day. Based on our observation, a similar pattern could be seen periodically each week.

The dataset consists of two sub-datasets. The content of both sub-datasets is identical, but one sub-dataset contains the real domain names, and the second contains the hashes of the domain names only. The hashes of domain names could be used for faster processing by machine learning models.

Each sub-dataset consists of 3 directories with single CSV files formed into a specific structure. The dataset scheme can be seen in Fig. 2. Each CSV file represents a single day. The detailed CSV file structure is described in Table 1 and described in detail in Table 2.

The individual CSV files consist of 1137 attributes, where the first attribute value corresponds to the observed IP address hidden behind the ID, and the following 104 attributes correspond to the daily record. The remaining 1032 attributes correspond to individual hourly records, where each hour consists of 43 attributes.

Due to various security reasons and the possibility of exploiting a given dataset for malicious purposes, the individual ip addresses were hidden behind index.

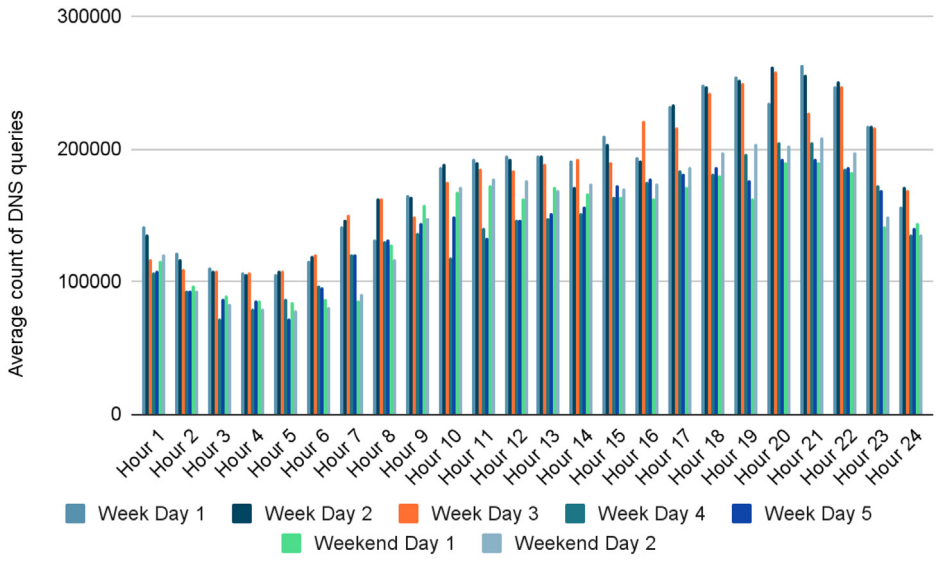

Fig. 1. The pattern of a typical weekly flow of DNS queries sorted according to the day of the week and hour of the day.

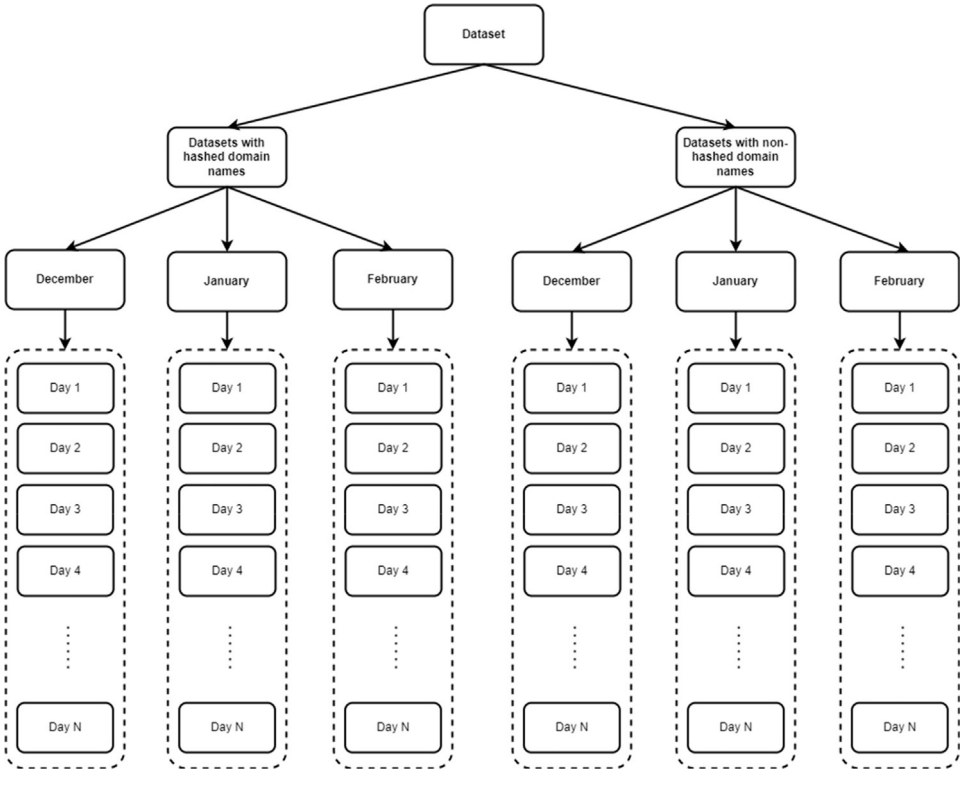

Fig. 2. User DNS Fingerprint Dataset structure, the records are stored per day in the directories representing single moths.

Table 1
General record structure.

Daily records			Hourly records		
Daily activity	Domain Categories	Ten most regularly visited domains during the day	Hour N (24 repetitions)		
			Overall hourly activity	Top ten most visited domains	The ten least visited domains

Table 2
Attributes of general record structures described more in detail.

Daily activity								
Number of queries	Number of unique domains	Number of types DNS query	Average daily activity	Average of the number of queries min and max h	Most active Hour (max h)	Least active hour (min h)	Number of queries in max h	Number of queries in min h

Domain Categories
... discussed in more detail in the text ...

Ten most regularly visited domains during the day (20 repeating columns)				
Domain Name/Hash	Number of hours	Domain Name/Hash	Number of hours	. . .

Overall hourly activity		
Number of queries	Number of unique domains	Number of DNS query types

Top ten most visited/least visited domains (20 repeating columns)				
Domain Name/Hash	Number of queries	Domain Name/Hash	Number of queries	. . .

3.1. Daily Records

The daily records consist of three sections that, with the total, describe the individual's behavior on the observed day. The individual sections are a summary of daily activity, visited domain categories, and a section of ten regularly visited domains. The daily records are spread across the first 105 attributes.

3.1.1. Daily Activity

Represents users daily activity over nine distinct attributes:

1. **Number of queries:** Total number of DNS queries performed in one day.
2. **Number of unique domains:** The total number of unique domains for which DNS queries were made in one day.

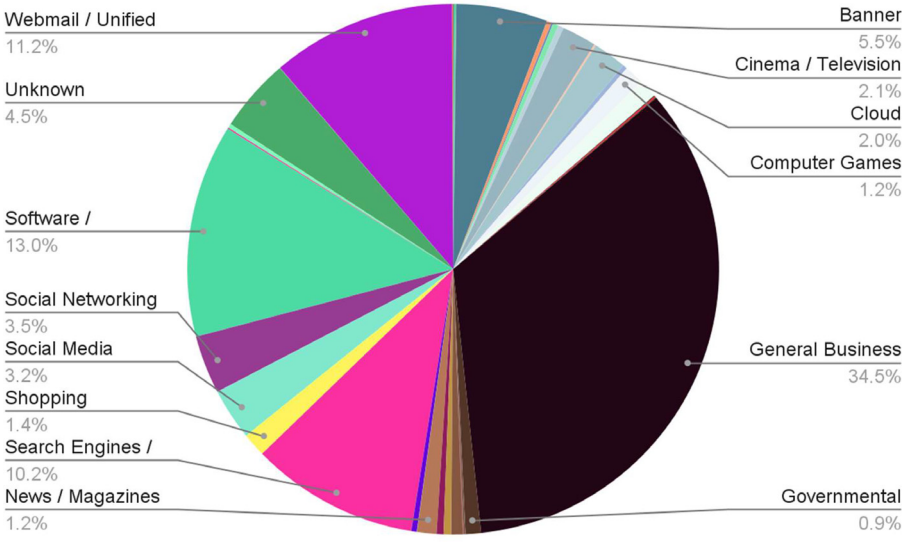

Fig. 3. Graphical representation of the domain categories access distribution for a selected day.

3. **Number of query types:** The total number of different types of DNS queries made in one day. These types can include, for example, A, AAAA, MX, etc.
4. **Average daily activity:** Average number of DNS queries per day.
5. **Most active hour (max h):** The hour with the most DNS queries performed in the day.
6. **Least active hour (min h):** The hour with the fewest DNS queries performed in the day.
7. **Number of queries in max h:** The total number of DNS queries made during the most active hour.
8. **Number of queries in min h:** The total number of DNS queries made during the least active hour.
9. **Average of the number of queries min and max h:** Average number of DNS queries during the most and least active hour of the day.

3.1.2. Categories

The following 75 attributes describe the number of visits to a domain with a specific content type. Individual domain types were annotated using IBM's x-force [10]. An example distribution of query categories during the day can be seen in Fig. 3.

1. **Abortion:** This category contains abortion-related websites and information, providing support and information for women in difficult situations.
2. **Alcohol:** Domains in this category are dedicated to alcohol, offering information about different types of drinks, tastings, and experiences related to alcohol consumption.
3. **Anonymisation Services:** These domains provide services to anonymise online activity, allowing users to protect their identity and privacy online.
4. **Architecture / Construction / Furniture:** The category includes websites related to architecture, construction, and furniture, providing design inspiration and information.
5. **Arts / Museums / Theatres:** Domains in this category are dedicated to art, museums, and theatres, presenting works of art and cultural events.
6. **Auctions / Classified Ads:** This category contains websites for auctions and classifieds, allowing users to buy and sell online.
7. **Banking:** Domains related to banking, providing information about banking services, loans, and financial transactions.

8. **Banner Advertisements:** This category includes banner advertising websites showcasing strategies and trends in online advertising.
9. **Blogs / Bulletin Boards:** Domains in this category belong to blogs and discussion forums that share users' opinions, advice, and experiences.
10. **Botnet Command and Control Server:** This category contains domains associated with botnet control servers that can be used for cyber attacks.
11. **Brokers / Stock Exchange:** Domains in this category are dedicated to brokers and stock exchanges, providing information on securities trading and financial investments.
12. **Business Networking:** The category contains domains related to business social networking, promoting communication and collaboration between businesses.
13. **Chat:** Domains providing online chat services, allowing users to communicate in real time with others.
14. **Cinema / Television:** This category includes domains dedicated to cinema and television, providing information about films, TV shows, and actors.
15. **Cities / Regions / Countries:** Domains in this category are dedicated to specific cities, regions, and countries, offering information about the culture and geography of the location.
16. **Cloud:** Domains associated with cloud technologies, providing information about cloud storage and online services.
17. **Communication Services:** This category contains domains providing various communication services, including email, telephony and video calling.
18. **Computer Crime / Hacking:** Domains related to cybercrime and hacking, providing information on security and prevention.
19. **Computer Games:** This category includes domains related to computer games, including reviews, news, and online gaming communities.
20. **Cryptocurrency Mining:** Domains related to cryptocurrency mining, providing information on mining farms and technologies.
21. **Dating:** This category contains domains focused on online dating, providing platforms for matchmaking and dating.
22. **Digital Postcards:** Domains associated with digital postcards, allowing users to share virtual greetings and pictures.
23. **Early Warning:** This category contains domains related to early warning systems, providing information on safety and prevention.
24. **Education:** Domains providing information about education, online courses and schools, supporting the educational process.
25. **Environment / Climate / Pets:** The category includes domains related to environment, climate and pets, providing information about caring for the planet and animals.
26. **Erotic / Sex:** Domains related to erotic or sexual content, providing adult material and sex education.
27. **Fashion / Cosmetics / Jewellery:** This category contains websites dedicated to fashion trends, cosmetics and jewellery, presenting news and style inspiration.
28. **Financial Services / Insurance / Real Estate:** Domains in this category are dedicated to financial services, insurance and real estate, offering information about investments and real estate offers.
29. **Gambling / Lottery:** This category contains domains related to gambling and lotteries, providing information on betting and odds of winning.
30. **General Business:** Domains related to general business, offering information about business, business strategies and trends.
31. **Governmental Organisations:** Domains in this category are dedicated to government organizations, providing information about government and public services.
32. **Health:** This category contains health-related websites offering information on treatment, prevention and wellness.
33. **Humour / Cartoons:** Domains related to humor and cartoons, presenting funny content and cartoons.

34. **Illegal Activities:** This category contains domains associated with illegal activities, including information about illegal trades and activities.
35. **Illegal Drugs:** Domains related to illegal drugs, providing information about drugs and their illegal trade.
36. **Instant Messaging:** Domains providing instant messaging services, allowing users to communicate quickly.
37. **IT Security / IT Information:** The category includes domains related to IT security and information technology, providing advice and updates.
38. **Job Search:** Domains in this category are dedicated to job search, offering platforms for job hunting and professional development.
39. **Literature / Books:** This category contains websites related to literature and books, offering reviews, book tips and literary events.
40. **Malware:** Malware-related domains, providing information about computer viruses, malware and their removal.
41. **Mobile Telephony:** This category contains domains related to mobile telephony, providing information about phones, reviews and news in the field of mobile technology.
42. **Music / Radio Broadcast:** Domains in this category are dedicated to music and radio broadcasting, providing information about artists, songs and music events.
43. **News / Magazines:** This category contains websites related to news and magazines, providing up-to-date information on events and topics.
44. **Non-Governmental Organisations:** Domains in this category are dedicated to non-profit organizations, providing information about charitable activities and volunteering.
45. **Personal Web Sites:** Domains focused on individuals' personal websites, showcasing their interests, ideas and lifestyle.
46. **Phishing URLs:** Domains associated with phishing, containing fake websites to obtain sensitive information from users.
47. **Platform as a Service:** The category includes domains related to Platform as a Service, providing information about cloud platforms.
48. **Political Extreme / Hate / Discrimination:** Domains associated with political extremism, hatred, and discrimination, containing controversial political material.
49. **Political Parties:** Domains dedicated to political parties, providing information on political party programmes and activities.
50. **Pornography:** This category contains domains with adult pornographic content, presenting erotic material.
51. **Recreational Facilities / Theme Parks:** Domains in this category are dedicated to recreational facilities and theme parks, offering information on fun activities.
52. **Religion:** This category contains websites related to religion, offering information about faith, religious traditions, and spiritual growth.
53. **Restaurants / Entertainment Venues:** Domains in this category are dedicated to restaurants and entertainment venues, providing information on menus, reviews, and cultural events.
54. **Search Engines / Web Catalogues / Portals:** This category contains domains related to search engines, web catalogues, and portals providing services to facilitate navigation on the Internet.
55. **Sects:** This category contains sites about sects, cults, occultism, Satanism etc..
56. **Self-Help / Addiction:** This category contains websites related to self-help and overcoming addiction, offering support and tips to improve your life.
57. **Shopping:** Domains in this category are dedicated to online shopping, providing information about products, discounts, and reviews.
58. **Social Media:** The category contains domains associated with social media, allowing users to share content and interact online.
59. **Social Networking:** Domains dedicated to social networking, encourage people to connect and share information.

60. **Software as a Service:** The category includes domains related to software as a service, providing information about cloud-based software solutions.
61. **Software / Hardware:** Domains in this category are dedicated to software and hardware, providing information on technology news and reviews.
62. **Spam URLs:** Domains associated with spam, containing unsolicited advertising and propaganda material.
63. **Sports:** This category contains sports-related websites, providing information about matches, players and sporting events.
64. **Swimwear / Lingerie:** Domains in this category are dedicated to swimwear and bras, offering information on the latest trends and styles.
65. **Tobacco:** The category contains tobacco-related domains, providing information on different types of tobacco and smoking supplies.
66. **Toys:** Toy related domains offering information on children's toys, reviews and selection tips.
67. **Travel:** This category contains travel related websites offering information on destinations, accommodation and travel experiences.
68. **Unknown:** Domains that cannot be clearly assigned to a specific category and are marked as unknown.
69. **Vehicles:** This category contains vehicle-related domains, providing information about cars, motorcycles and vehicles.
70. **Violence / Extreme:** Domains associated with violence and extremism, containing controversial material.
71. **Warez / Software Piracy:** This category contains domains associated with illegal software downloads and piracy.
72. **Weapons / Military:** Domains dedicated to weapons and military topics, providing information on weapons and military history.
73. **Webmail / Unified Messaging:** This category contains domains related to webmail and unified messaging services, enabling users to manage their email effectively.
74. **Web Site Translation:** Domains providing website translation services, facilitating access to content in different languages.
75. **Web Storage:** This category contains domains associated with web storage, offering online data storage and sharing options.

3.1.3. Regularly Visited Domains During the Day

The following 20 attributes represent the 10 domains visited in most hours of the day. The structure consists of ten pairs, where each pair contains the domain name or its hash of the name and the corresponding value. Let the index of the domain be k, $k \in \langle 0, 9 \rangle$. The detailed structure of the pair record is as follows:

1. ***k* most regularly visited domain:** Indicates the name or hash of the name of the k-th most regularly visited domain in a day.
2. **Occurrence of *k* most regularly visited domain:** Represents the total number of hours the domain was seen.

3.2. Hourly Records

The hourly records describe the user activity within a specific hour. The hourly records consist of three sections - hourly activity, most visited domains, and least visited domains. The hourly records represent an entire day and consist of a total of 1032 attributes. Each hour corresponds to 43 attributes. The hourly record is repeated 24 times. Here is a description of one hourly record, let's call the index of an hour i, $i \in \langle 0, 23 \rangle$.

Table 3
Structure of the record in file domain_hash_category.CSV.

Domain name	Hash of Domain name	Category of Domain name	No. Category

3.2.1. Overall Hourly Activity

This section represents the total activity of the user during a specific hour. The following three attributes represent it:

1. **Hour *i* number of queries:** Indicates the total number of DNS queries performed during one hour.
2. **Hour *i* number of unique domains:** Represents the total number of unique domains for which DNS queries were made during one hour.
3. **Hour *i* number of query types:** Indicates the total number of different types of DNS queries that were performed in one hour. These types can include, for example, A, AAAA, MX, etc.

3.2.2. Most Visited Domains

The structure of the record of the most visited domain consists of 20 attributes. Each pair of attributes corresponds to the domain name or hash of the domain name and the number of queries in an hour. The record contains the first 10 most visited domains; let's call the index of the domain k, $k \in <0, 9>$.

Here is a brief description of attributes building a pair for one domain:

1. **Hour *i* most visited domain *k*:** The identifier of the kth most visited domain in a given hour. It can be either name or hash.
2. **Hour *i* most visited domain *k* count:** The number of DNS queries made to the k-th most visited domain in a given hour.

This record is repeated 10 times for each most visited domain.

3.2.3. Least Visited Domains

The structure of the least visited domains corresponds to the structure and format of the most visited domains, except that the least visited domains are tracked instead of the most visited domains. This information allows tracking the least active part of DNS traffic in a given period.

The dataset contains in the folder "hashed_domain_names" the file called "domain_hash_category.csv". This file contains records with the structure described in Table 3. The record consist of domain name, hash of the domain name, and the category. The category represents the type of content related to the domain name.

4. Experimental Design, Materials, and Methods

This section describes how we gathered the data and used the environment. The data collection topology is depicted in Fig. 4. The users of an Internet Service Provider Network forwarded DNS queries to a private DNS server. The DNS server stored the sent queries in a log file with a similar structure as described below. The concretely used DNS server tool was the BIND 9 [11], which is currently one of the most used DNS server solutions worldwide. The log files of BIND 9 were subsequently transformed into CSV files described in the previous section.

A record stored in the BIND 9 DNS Server log file looked as follows:

```
01-Jan-2024 00:00:00.001 client 10.17.6.12#39646 (redirector.googlevideo.com): query: redirector.googlevideo.com IN A + (10.0.1.1)
```

1. **Time stamp:** 01-Jan-2024 00:00:00.001
2. **Network user:** client 10.17.6.12#39646

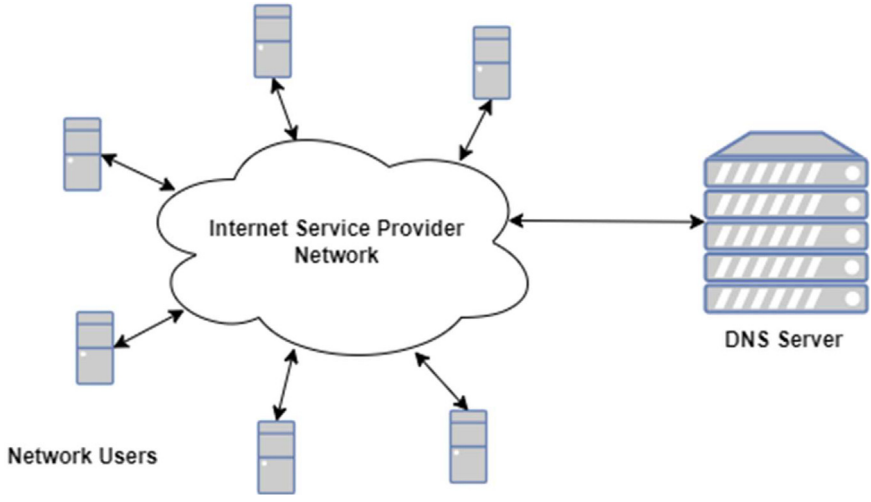

Fig. 4. Internet Service Provider Network containing the users and a DNS Server. The DNS server was used by the network users only.

3. **The Record:** (redirector.googlevideo.com) query: redirector.googlevideo.com IN A + (10.0.1.1)

Limitations

None.

Ethics Statement

Our work does not involve studies with animals and humans.

Data Availability

A User DNS Fingerprint Dataset (Original data) (Zenodo).

CRediT Author Statement

Josef Zápotocký: Data curation, Software, Writing – original draft; **Jan Fiala:** Visualization; **Jan Fesl:** Conceptualization, Methodology, Writing – original draft, Supervision.

Acknowledgments

The authors would also like to acknowledge Czech Technical in Prague, Faculty of Information Technology, and the University of South Bohemia, Faculty of Science, for providing the technical stuff necessary for the data set creation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Zápotocký, J. Fiala, J. Fesl, A user DNS fingerprint dataset, available on-line. 2024, https://zenodo.org/records/10887463 (accessed March 25, 2024).
[2] M. Laštovička, M. Husák, P. Velan, T. Jirsík, P. Čeleda, Passive operating system fingerprinting revisited: Evaluation and current challenges, Comput. Netw. 229 (2023) 109782.
[3] D. Herrmann, M. Kirchler, J. Lindemann, M. Kloft, Behavior-based tracking of internet users with semi-supervised learning, in: Proceedings of the 2016 14th Annual Conference on Privacy, Security and Trust (PST), Auckland, New Zealand, 2016, pp. 596–599, doi:10.1109/PST.2016.7906992.
[4] D.W. Kim, J. Zhang, Deriving and measuring DNS-based fingerprints, J.Inf. Secur. Appl. 36 (2017) 32–42.
[5] M. Singh, M. Singh, S. Kaur, Detecting bot-infected machines using DNS fingerprinting, Digit. Investig. 28 (2019) 14–33 De.
[6] T. Matsunaka, A. Yamada, A. Kubota, Passive OS fingerprinting by DNS traffic analysis, in: Proceedings of the 2013 IEEE 27th International Conference on Advanced Information Networking and Applications (AINA), Barcelona, Spain, 2013, pp. 243–250, doi:10.1109/AINA.2013.119.
[7] D. Herrmann, C. Banse, H. Federrath, Behavior-based tracking: exploiting characteristic patterns in DNS traffic, Comput. Secur. 39A (2013) 17–33.
[8] Z. Jia, Z. Han, Research and analysis of user behavior fingerprint on security situational awareness based on DNS log, in: Proceedings of the 2019 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC), Beijing, China, 2019, pp. 1–4, doi:10.1109/BESC48373.2019.8963120.
[9] Q. Lai, C. Zhou, H. Ma, Z. Wu, S. Chen, Visualizing and characterizing DNS lookup behaviors via log-mining, Neurocomputing. 169 (2015) 100–109.
[10] IBM: 2021 X-force threat intelligence index, Netw. Secur. 2021 (Issue 3) (2021).
[11] Internet system consortium, Bind 9, 2024, available on-line, https://www.isc.org/bind/ (accessed March 25, 2024).