Geedge Cases: Censorship Measurement Insights from the Geedge Networks Leak

Jade Sheffey
University of Massachusetts Amherst
Amherst, Massachusetts, USA
jsheffey@cs.umass.edu

Ali Zohaib
University of Massachusetts Amherst
Amherst, Massachusetts, USA
azohaib@umass.edu

Mingshi Wu
GFW Report
gfw.report@protonmail.com

Amir Houmansadr
University of Massachusetts Amherst
Amherst, Massachusetts, USA
amir@cs.umass.edu

Abstract

Geedge Networks is a network security company that builds Internet censorship software for both China and foreign authoritarian regimes. The September 2025 Geedge Networks leak (GNL) exposed 572 GiB of internal documents, source code, and binaries from Geedge Networks and the related MESA lab. We analyze 6,915,266 domains extracted from the GNL and compare them against the two most widely used domain lists in censorship research: Tranco and the CitizenLab test lists.

Our analysis across 5 locations reveals that 298,955 censored GNL domains (93.7% of all censored GNL domains) are not included in either Tranco or the CitizenLab lists. While Tranco captures globally popular sites and CitizenLab monitors sensitive content categories, the GNL provides a vendor-side perspective on which domains commercial censorship systems consider of interest. By correlating censored domains with filepaths in the GNL, we reveal files containing domain lists likely to be related to censorship done by customers of Geedge Networks.

Keywords

Censorship measurement, Censored domains, Geedge Networks

1 Introduction

Understanding what content authoritarian regimes monitor is fundamental to censorship research. Researchers typically rely on either lists of popular domains (Tranco [14], historically Alexa [1]) or curated domain lists containing content historically or likely to be censored based on controversial content (Citizen Lab test lists [3]). The Geedge Networks leak (GNL), released 11 September 2025 [8] provides unprecedented ground truth: 572 GiBof internal documents from a Chinese company that sells network censorship software both internally to China and to other countries including Kazakhstan, Pakistan, Myanmar, and Ethiopia [2]

The GNL represents the first opportunity to examine what domains a commercial censorship vendor actually configures their deep packet inspection (DPI) systems to monitor. Unlike academic censorship measurements that test popular or controversial domains, these leaked files reveal operational priorities: what governments *pay* to block.

In this work, we set out to analyze domains found in the GNL to answer two questions:

(1) Which domains are of interest to Geedge Networks or its customers?
(2) Which components of the GNL are most relevant to global censorship?

Our analysis makes three key contributions:
Domain dataset: We extract and analyze 6,915,266 domains from the GNL, representing actual commercial censorship targets in four countries.
Systematic comparison: We systematically compare the censorship of domains found in the GNL with the Tranco top 1 million list and Citizen Lab's test lists (global, China, Myanmar, Pakistan, and Algeria).
Attribution: We correlate files in the GNL with censorship measurements to identify important documents relevant to censorship research.

2 Background and Related Work

2.1 The Geedge Networks Leak

Geedge Networks, founded in 2018 by Fang Binxing (known as the "Father of the Great Firewall" [25]), commercializes and exports Internet censorship infrastructure. On 11 September 2025, the hacktivist collective Enlace Hacktivista released to the general public 572 GiBof internal documents from Geedge Networks and the related MESA lab including source code repositories, project management records, client deployment configurations, and operational data [8]. The leak revealed deployments in at least five countries, including Kazakhstan, Pakistan, Myanmar, Ethiopia, and China [2, 7, 12, 13].

The GNL is composed of five main components, listed in Table 1. Key components of the GNL include source code (`mesalab_git`), documentation (`geedge_docs`, `geedge_jira`, `mesalab_docs`, `misc`), and binary packages (`mirror`).

2.2 Domain Lists in Censorship Research

Censorship measurement studies require selecting which domains to test. Three approaches dominate:

This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license visit https://creativecommons.org/licenses/by/4.0/ or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.
Free and Open Communications on the Internet 2025(X), 1–5
© 2025 Copyright held by the owner/author(s).
https://doi.org/XXXXXXX.XXXXXXX

Table 1: Overview of the structure of the GNL. Size is before preprocessing while file count listed is after preprocessing (§ 3.1)

Component	Description	Size	Files
`mirror`	Red Hat Linux RPM repository	463 GiB	59507
`geedge_docs`	Internal Geedge documentation	14 GiB	73900
`geedge_jira`	Internal Geedge JIRA issue tracker	2.6 GiB	6128
`mesalab_docs`	Internal MESA lab documentation	33 GiB	28081
`mesalab_git`	Git repositories for MESA lab	60 GiB	236292
`misc`	Misc DOCX files in the root folder	1.9 MiB	14

Popularity-based lists such as Tranco [14] aggregate multiple data sources (Chrome User Experience Report, Cloudflare Radar, Cisco Umbrella, Majestic Million, Farsight Security) over 30-day windows to produce stable rankings of 1M domains. Popularity lists are often used in Internet censorship research to measure censorship of mainstream content and quantify the breadth of filtering.
Curated sensitive content such as Citizen Lab's test lists [3] provide hundreds of curated URLs per country across 30 standardized categories (political criticism, human rights, LGBT, religion, news media, etc.). Created by volunteers, these lists focus on content likely to be censored, and are used by OONI [6], Censored Planet [17], and ICLab [15].
Automated discovery approaches like GFWatch [10] or GFWeb [9] use top-level domain (TLD) zone data from ICANN [11]. However, a list of this scale may trigger defensive measures from censors [10]. Additionally, TLD zone data only covers second-level domains (SLDs). GFWeb [9] additionally uses FQDNs from the Common Crawl [4] dataset.

No prior work has compared the censorship of domains in major domain lists with domains extracted from the leaked data of a commercial censorship vendor.

3 Methodology

3.1 Preprocessing

We performed our analysis on the GNL as uploaded by Enlace Hacktivista [8], with the following preprocessing steps:

- `geedge_docs.tar.zst`, `geedge_jira.tar.zst`, `mesalab_docs.tar.zst`, `mesalab_git.tar.zst`, `mirror/repo.tar` were extracted.
- Git repository bundles in `mesalab_git` were cloned to allow accessing the files within
- Tesseract [19–23] was used to extract text from all files with MIME type `image/*` to a separate folder.

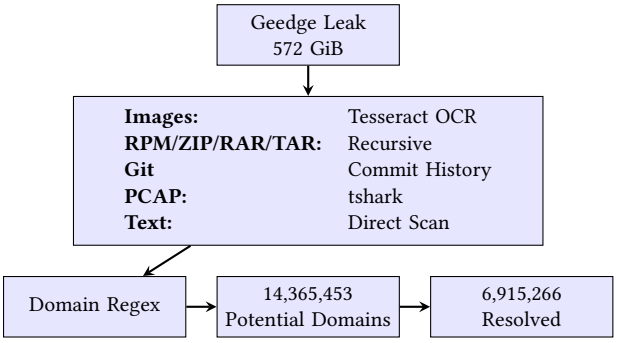

Figure 1: Domain extraction pipeline. Each file was matched to a type, with special extraction procedures for certain common file formats. After extraction, all files are scanned using a regular expression, filtered, deduplicated, and tested using multiple resolvers.

3.2 Domain Extraction

Figure 1 shows our domain extraction pipeline. First, the file's MIME type is determined using its extension, with the `magic` [18] library used as a fallback. Next, each file is handled as either raw binary data, or processed using a specialized handler. Archives (RPM, zip, JAR, TAR) are processed recursively, Git repositories have both their current state and commit diffs scanned, and PCAP/PCAPNG files are processed using `tshark` [24] in PDML mode.

Once binary data has been extracted from each file, we use a regular expression with TLDs from Mozilla's public suffix list [16] to extract all domains. The regular expression is:

```
((a-zA-Z0-9\-_)+\.)+(com|net|<TLDs>)\.?
```

Additionally, we include parents of found domains, up to the SLD. Using this strategy, we produce a list of 14,365,453 potential domains. To remove strings that may be valid domains but in practice are not (e.g. `libstdcpp.so`), we send A and AAAA DNS requests for each domain to `1.1.1.1` from an uncensored network. After resolution filtering, we produce a final list of 6,915,266 domains. Additionally, we make HTTP and HTTPS requests to each resolved domain to establish a baseline.

3.3 Lists

We compare our GNL domain list against two well-known benchmark lists: **Tranco**[1] [14] and the **Citizen Lab test lists**[2] [3].

3.4 Measurements

To measure the relevance of the GNL domain list to global censorship, we perform measurements across vantage points in China (Guangzhou and Nanjing), Myanmar, Pakistan, and Algeria. While Kazakhstan and Ethiopia are known customers of Geedge Networks [12], we were unable to obtain access to appropriate vantage points in these countries. We perform the same measurements using the Tranco and Citizen Lab lists. For Chinese vantage points, we measure censorship using injected DNS responses, analyzing PCAP files to identify domains that receive hijacked responses rather than

[1] Generated Oct 14, 2025. Available at https://tranco-list.eu/list/3Q3XL
[2] Revision 1a725026abe96db06116f33b2c88b27727cac002

legitimate DNS resolution. For Pakistan, Myanmar, and Algeria, we test for SNI-based censorship by establishing TLS connections and detecting early connection termination (EOF) during the TLS handshake, which indicates SNI-based blocking of domain names. We use DNS-based measurements for China because DNS censorship is highly reliable—queries to non-existent IP addresses still receive injected responses, so any response definitively indicates censorship. TLS-based measurements are used elsewhere as they detect censorship through connection failures, which are inherently less reliable indicators. For both DNS and TLS measurements, we perform 25 tests per domain to ensure intermittently blocked domains are included.

4 Results

4.1 Measurements

Table 2 shows censorship measurements from four countries using different domain lists. DNS measurements from Chinese vantage points (Guangzhou and Nanjing) show that GNL domains account for around 25% of censored CitizenLab domains and around 50% of censored Tranco domains. TLS measurements from Algeria, Myanmar, and Pakistan reveal additional censorship patterns using SNI-based blocking detection.

The "Unique GNL" column reveals that 298,955 domains (93.7% of all GNL censored domains) are *experimentally verified as censored* through our DNS and TLS measurements, yet appear in neither Tranco nor CitizenLab test lists. Importantly, we do not assume domains are censored simply because they appear in the GNL—explicit blocklists are rare in the leak, as they constitute sensitive customer data. Instead, domains in the GNL represent content "of interest" to Geedge or its customers: either registered in Geedge systems or observed via network monitoring. Our approach extracts these domains of interest, then uses real-world measurements to identify which are actually censored, providing coverage that complements current academic measurement methodologies.

4.2 Domain Sources

By correlating experimentally-verified censored domains with their source files in the GNL, we can identify which files are most likely related to actual censorship operations rather than internal research or testing. Our domain extraction pipeline generates detailed logs of which files contain which domains, enabling us to trace censored domains back to their origins within Geedge's internal systems.

Table 3 presents the most significant files containing censored domains, organized by geographic region and business context. We have filtered out common domain lists (Alexa [1], SecRank, Tranco) that appear to be used for internal research rather than active censorship deployment.

We make the following observations based on frequent sources:

- **SNI-based surveillance**: The largest number of domains comes from MESA lab's specialized SNI datasets (E21-SNI-Top200w.txt, E21-SNI-Top120W, etc). These datasets do not seem to come from popular domain lists, and instead appear to be gathered from network taps.
- **VPN infrastructure mapping**: Multiple files specifically target VPN and circumvention tools, including comprehensive NordVPN server lists and Psiphon CDN domains.

Of particular interest to censorship measurement research is `geedge_docs/TSGEN/attachments/48048462_attachments_白名单网站.txt`, where 白名单网站 means "whitelisted websites". This file is referenced by `geedge_docs/TSGEN/2021-10-24.html`, which describes how Geedge software is deployed in practice on a mobile telecom network in Quanzhou. The document describes allow rules, which include whitelisted domains, but also deny rules, which include blocked domains, fraudulent apps, user agents associated with fraud or prostitution, gambling domains, and even domains to intercept APK downloads. The interception of APK downloads is potentially related to Geedge's Appsketch program, which is used to reverse engineer domains, IP addresses, and characteristics of VPN apps for blocking.

5 Discussions

The Geedge Networks leak provides an unprecedented ground truth for understanding what commercial censorship vendors actually monitor. By comparing 6,915,266 extracted domains from the leak against the Tranco and Citizen Lab test lists, we find that **the GNL contains 298,955 censored domains not included in either standard test list**, providing a complementary source of censor-relevant domains.

Our analysis across 5 countries reveals three key findings: First, 93.7% of censored GNL domains (298,955 domains) appear in neither Tranco nor CitizenLab lists, indicating that commercial censorship vendors monitor a substantially different set of domains than those captured by popularity-based rankings or curated test lists. Second, while there is overlap between GNL and existing lists—the GNL captures between 37.1% and 61.6% of censored Tranco domains—each list makes a distinct contribution: Tranco informs which globally popular sites are censored, CitizenLab monitors types of sensitive content over time, and the GNL reveals which domains commercial vendors consider of interest. Third, commercial censorship vendors maintain sophisticated threat intelligence systems with over 57,000 domains in single monitoring datasets, far exceeding the scale of academic test lists. These findings suggest that **incorporating domains from commercial leak datasets** can complement existing methodologies by providing a vendor-side perspective on domain-based censorship.

Limitations: In general, censorship "rule lists" are somewhat rare in the GNL, as they are considered customer data. The domains we extract mainly seem to come from internal discussions of customer environments. We find that censorship researchers, much like anticensorship researchers, use lists of popular domains for their own internal research, creating some overlap. The GNL contains copies of popularity lists (Alexa, SecRank) that appear to be used for MESA lab research projects rather than deployed Geedge products; these are included in our 6.9M domain count. However, the 298,955 unique censored domains are by definition not in Tranco or CitizenLab, so this overlap does not affect our main finding. Additionally, our extraction process cannot currently process PDFs.

Ethical Considerations: While this research analyzes data obtained from a leak, we believe the broad public interest of this information overrides potential intellectual property concerns. Geedge Networks has enabled human rights violations around the world

Table 2: Censorship measurements comparing domain lists across vantage points. GNL Coverage denotes the percent of censored domains in a list covered by the GNL. Unique GNL is the count of censored GNL domains not in any other list.

Location	List	List Size	Censored	GNL Coverage	Unique GNL
Guangzhou/Nanjing	Citizen Lab Combined	37,919	2,696 (7.1%)	692 (25.7%)	-
Guangzhou/Nanjing	Citizen Lab China	589	243 (41.3%)	72 (29.6%)	-
Guangzhou/Nanjing	Tranco	1,000,000	7,821 (0.8%)	3,876 (49.6%)	-
Guangzhou/Nanjing	GNL	6,915,266	218,339 (3.2%)	-	211,746 (97.0%)
Algeria	Citizen Lab Combined	37,919	71 (0.2%)	36 (50.7%)	-
Algeria	Citizen Lab Algeria	403	22 (5.5%)	10 (45.5%)	-
Algeria	Tranco	1,000,000	86 (0.0%)	53 (61.6%)	-
Algeria	GNL	6,915,266	299 (0.0%)	-	198 (66.2%)
Myanmar	Citizen Lab Combined	37,919	109 (0.3%)	43 (39.4%)	-
Myanmar	Citizen Lab Myanmar	875	20 (2.3%)	3 (15.0%)	-
Myanmar	Tranco	1,000,000	1,713 (0.2%)	672 (39.2%)	-
Myanmar	GNL	6,915,266	3,131 (0.0%)	-	2,988 (95.4%)
Pakistan	Citizen Lab Combined	37,919	617 (1.6%)	221 (35.8%)	-
Pakistan	Citizen Lab Pakistan	670	28 (4.2%)	9 (32.1%)	-
Pakistan	Tranco	1,000,000	19,406 (1.9%)	7,209 (37.1%)	-
Pakistan	GNL	6,915,266	113,796 (1.6%)	-	98,992 (87.0%)
Total Unique GNL Domains:					**298,955**

Table 3: GNL files containing many censored domains.

Location	Count	Path	Description
Common	57,362	`mesalab_git/galaxy/.../entity_dataset/E21-SNI-Top200w.txt`	E21=Ethiopia [12]
Common	36,467	`mesalab_git/galaxy/.../entity_dataset/E21-SNI-Top120W-20221020.txt`	E21=Ethiopia [12]
Common	24,219	`mesalab_git/tsg/tsg-deploy/.../porn.csv`	Adult websites
Common	13,604	`mesalab_git/galaxy/.../entity_dataset/XJ-CUCC-SNI-Top200w.txt`	XJ=Xinjiang? [12]
Common	10,163	`mesalab_git/tango/maat/test/tsgrule/TSG_OBJ_FQDN.E21`	E21=Ethiopia [12]
China	7,016	`mesalab_git/intelligence-learning-engine/vpn-finder-plugins`	VPN host discovery
China	4,810	`geedge_docs/.../Nord VPN server List.txt`	NordVPN servers
China	475	`geedge_docs/TSGEN/attachments/48056407...20211025.txt`	Quanzhou block/allowlists
Myanmar	27	`geedge_docs/TSGEN/M22-VPN List.html`	M22=Myanmar [12]
Pakistan	68	`geedge_docs/TSGEN/.../Psiphon-CDN_20240430.json`	Psiphon domains
Algeria	11	`mesalab_docs/shu/.../mail.alakhbar.press.ma`	Moroccan mail servers

with its software, and its inner workings are of public interest. Additionally, the leaked data has already been provided publicly via multiple sources including Enlace Hactivista[8] and Distributed Denial of Secrets [5].

Data availability: The extraction code and full list of domains found in this work will be released publicly.

5.1 Future Work

While this work focuses on domain censorship specifically, the GNL reveals a massively developed censorship apparatus. Promising areas for future work include searching for IP addresses in the GNL, deeper analysis of documents such as PDFs, and improved OCR. Additionally, the Tranco and Citizen Lab lists are limited compared to larger censorship measurement lists such as Common Crawl or ICANN CZDS, which may offer even more overlap with the GNL domain list. Future analysis should also examine topic patterns among censored domains not found in standard lists, and measure domain transience to understand how many identified domains remain active over time.

Acknowledgments

This work was made possible by the anonymous source that released the GNL to the public. The work was supported in part by the NSF grant CNS-2333965, and by the Young Faculty Award program of the Defense Advanced Research Projects Agency (DARPA) under the grant DARPA-RA-21-03-09-YFA9-FP003. The views, opinions, and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References

[1] Amazon. 2022. Alexa - Top Sites. https://web.archive.org/web/20220226063254/https://www.alexa.com/topsites.

[2] Amnesty International. 2025. Pakistan: Shadows of Control: Censorship and Mass Surveillance in Pakistan. https://www.amnesty.org/en/documents/asa33/0206/2025/en/.

[3] Citizen Lab. 2025. Citizenlab/Test-Lists. https://github.com/citizenlab/test-lists.

[4] Common Crawl. 2025. Common Crawl - Open Repository of Web Crawl Data. https://commoncrawl.org.

[5] DDOSecrets. 2025. Geedge Networks - Distributed Denial of Secrets. https://ddosecrets.com/article/geedge-networks.

[6] Arturo Filastò and Jacob Appelbaum. 2012. OONI: Open Observatory of Network Interference. In *Free and Open Communications on the Internet*. USENIX. https://www.usenix.org/system/files/conference/foci12/foci12-final12.pdf

[7] James Griffiths. 2025. Leaked Files Show a Chinese Company Is Exporting the Great Firewall's Censorship Technology. https://www.theglobeandmail.com/world/article-leaked-files-show-a-chinese-company-is-exporting-the-great-firewalls/. *The Globe and Mail* (8 Sept. 2025).

[8] Enlace Hactivista. 2025. Geedge Networks - Enlace Hacktivista. https://enlacehacktivista.org/index.php/Geedge_Networks.

[9] Nguyen Phong Hoang, Jakub Dalek, Masashi Crete-Nishihata, Nicolas Christin, Vinod Yegneswaran, Michalis Polychronakis, and Nick Feamster. 2024. GFWeb: Measuring the Great Firewall's Web Censorship at Scale. In *USENIX Security Symposium*. USENIX. https://www.usenix.org/system/files/sec24fall-prepub-310-hoang.pdf

[10] Nguyen Phong Hoang, Arian Akhavan Niaki, Jakub Dalek, Jeffrey Knockel, Pellaeon Lin, Bill Marczak, Masashi Crete-Nishihata, Phillipa Gill, and Michalis Polychronakis. 2021. How Great is the Great Firewall? Measuring China's DNS Censorship. In *USENIX Security Symposium*. USENIX. https://www.usenix.org/system/files/sec21-hoang.pdf

[11] ICANN. 2025. Centralized Zone Data Service. https://czds.icann.org/help.

[12] InterSecLab. 2025. *The Internet Coup: A Technical Analysis on How a Chinese Company is Exporting The Great Firewall to Autocratic Regimes*. Technical Report. InterSecLab. https://interseclab.org/research/the-internet-coup/

[13] Justice For Myanmar. 2025. *Silk Road of Surveillance: The role of China's Geedge Networks and Myanmar telecommunications operators in the junta's digital terror campaign*. Technical Report. Justice For Myanmar, Myanmar. https://www.justiceformyanmar.org/stories/silk-road-of-surveillance

[14] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. 2019. Tranco: A Research-Oriented Top Sites Ranking Hardened against Manipulation. In *Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS 2019)*. NDSS, San Diego, California., 1–15. https://doi.org/10.14722/ndss.2019.23386

[15] Arian Akhavan Niaki, Shinyoung Cho, Zachary Weinberg, Nguyen Phong Hoang, Abbas Razaghpanah, Nicolas Christin, and Phillipa Gill. 2020. ICLab: A Global, Longitudinal Internet Censorship Measurement Platform. In *Symposium on Security & Privacy*. IEEE. https://people.cs.umass.edu/~phillipa/papers/oakland2020.pdf

[16] Public Suffix List. 2025. Publicsuffix/List. https://github.com/publicsuffix/list.

[17] Ram Sundara Raman, Prerana Shenoy, Katharina Kohls, and Roya Ensafi. 2020. Censored Planet: An Internet-wide, Longitudinal Censorship Observatory. In *Computer and Communications Security*. ACM. https://www.ramakrishnansr.com/assets/censoredplanet.pdf

[18] Robo9k. 2025. Robo9k/Rust-Magic. https://github.com/robo9k/rust-magic.

[19] Faisal Shafait and Ray Smith. 2010. Table detection in heterogeneous documents.. In *Document Analysis Systems* (2010-07-07) *(ACM International Conference Proceeding Series)*, David S. Doermann, Venu Govindaraju, Daniel P. Lopresti, and Premkumar Natarajan (Eds.). ACM, 65–72. http://dblp.uni-trier.de/db/conf/das/das2010.html#ShafaitS10

[20] Ray Smith. 2007. An Overview of the Tesseract OCR Engine. In *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition*. IEEE Computer Society, Washington, DC, USA, 629–633. https://storage.googleapis.com/pub-tools-public-publication-data/pdf/33418.pdf

[21] Ray Smith. 2009. Hybrid Page Layout Analysis via Tab-Stop Detection. In *ICDAR '09: Proceedings of the 2009 10th International Conference on Document Analysis and Recognition*. IEEE Computer Society, Washington, DC, USA, 241–245. https://doi.org/10.1109/ICDAR.2009.257

[22] Ray Smith, Daria Antonova, and Dar-Shyang Lee. 2009. Adapting the Tesseract Open Source OCR Engine for Multilingual OCR.. In *MOCR '09: Proceedings of the International Workshop on Multilingual OCR* (Barcelona, Spain, 2009-07-25) *(ACM International Conference Proceeding Series)*, Venu Govindaraju, Premkumar Natarajan, Santanu Chaudhury, and Daniel P. Lopresti (Eds.). ACM, 1–8. https://doi.org/10.1145/1577802.1577804

[23] Ranjith Unnikrishnan and Ray Smith. 2009. Combined Orientation and Script Detection using the Tesseract OCR Engine. In *MOCR '09: Proceedings of the International Workshop on Multilingual OCR* (Barcelona, Spain), Venu Govindaraju, Premkumar Natarajan, Santanu Chaudhury, and Daniel P. Lopresti (Eds.). ACM, New York, NY, USA, 1–7. https://doi.org/10.1145/1577802.1577809

[24] Wireshark. 2025. D.2. Tshark: Terminal-based Wireshark. https://www.wireshark.org/docs/wsug_html_chunked/AppToolstshark.html.

[25] Mingshi Wu. 2025. Geedge & MESA Leak: Analyzing the Great Firewall's Largest Document Leak. https://gfw.report/blog/geedge_and_mesa_leak/en/.