Geedge Cases

Censorship Measurement Insights from the Geedge Networks Leak

Jade Sheffey, Ali Zohaib, Mingshi Wu, Amir Houmansadr

University of Massachusetts Amherst, GFW Report

19 Feb 2026

The Problem: What Do Censors Actually Block?

Current approaches:

- **Popularity lists** (Tranco): Top 1M domains
- **Curated lists** (CitizenLab): Sensitive content categories and user reports
- **Automated discovery**: TLD zone data, Common Crawl

The gap:

- We test what we *think* censors block
- No ground truth from censor's perspective
- We may be missing out on niche domains

The Problem: What Do Censors Actually Block?

Current approaches:

- **Popularity lists** (Tranco): Top 1M domains
- **Curated lists** (CitizenLab): Sensitive content categories and user reports
- **Automated discovery**: TLD zone data, Common Crawl

The gap:

- We test what we *think* censors block
- No ground truth from censor's perspective
- We may be missing out on niche domains

It would be great if we could see inside a censor's systems.

The Geedge Networks Leak (GNL)

Geedge Networks:

- Founded by Fang Binxing (“Father of the Great Firewall”)
- Builds censorship infrastructure for both China and internationally
- Customers: China, Kazakhstan, Pakistan, Myanmar, Ethiopia

The Geedge Networks Leak (GNL)

Geedge Networks:

- Founded by Fang Binxing ("Father of the Great Firewall")
- Builds censorship infrastructure for both China and internationally
- Customers: China, Kazakhstan, Pakistan, Myanmar, Ethiopia

September 2025: Anonymous source releases **572 GiB** of internal data from Geedge Networks

Component	Size
`mirror` (RPM repo)	463 GiB
`geedge_docs`	14 GiB
`mesalab_git`	60 GiB
`mesalab_docs`	33 GiB
`geedge_jira`	2.6 GiB

The Geedge Networks Leak (GNL)

Geedge Networks:

- Founded by Fang Binxing ("Father of the Great Firewall")
- Builds censorship infrastructure for both China and internationally
- Customers: China, Kazakhstan, Pakistan, Myanmar, Ethiopia

September 2025: Anonymous source releases **572 GiB** of internal data from Geedge Networks

Component	Size
`mirror` (RPM repo)	463 GiB
`geedge_docs`	14 GiB
`mesalab_git`	60 GiB
`mesalab_docs`	33 GiB
`geedge_jira`	2.6 GiB

Are there censorship rules in the GNL?

Research Questions

RQ1: Which domains are of interest to Geedge Networks or its customers?

Research Questions

RQ1: Which domains are of interest to Geedge Networks or its customers?

RQ2: Do domains in the GNL actually get censored in practice, and if so, which files contain censored domains?

What's in the Leak?

Source Code (`mesalab_git` - 60 GiB):

- 236,292 files across Git repositories
- DPI engine
- Deployment configuration

Documentation (`geedge_docs` - 14 GiB):

- 73,900 internal documents
- Customer deployment guides

JIRA (`geedge_jira` - 2.6 GiB):

- Issue tracking and project management

Binary Packages (`mirror` - 463 GiB):

- 59,507 RPM packages
- Compiled deployment artifacts

MESA Lab Research (`mesalab_docs` - 33 GiB):

- 28,081 research documents
- Network capture files (PCAP)
- Other miscellaneous datasets

The Challenge: Finding Domains

Domains are scattered across many formats:

Text-based:

- Source code (`.py`, `.java`, `.go`)
- Configuration files (`.json`, `.yaml`)
- Documentation (`.html`, `.txt`, `.md`)
- Deployment scripts

Binary:

- RPM packages (nested archives)
- Compiled binaries
- Database dumps

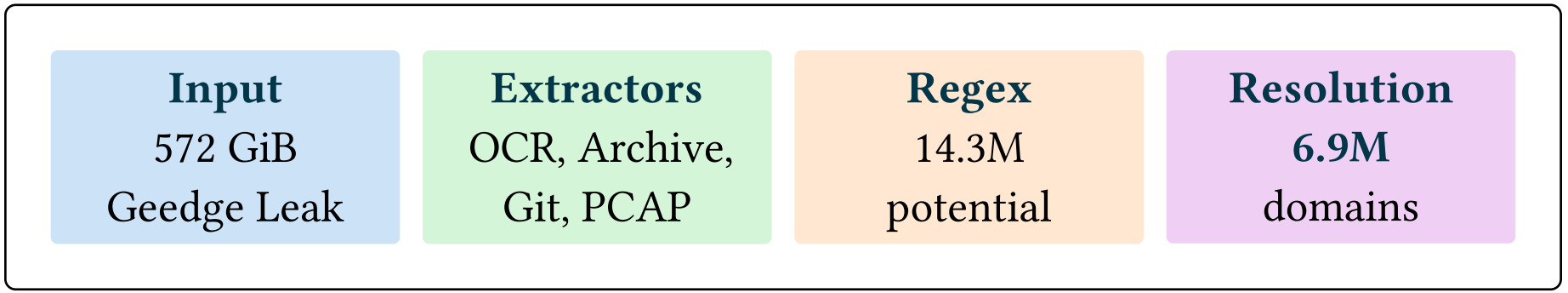
Specialized:

- Git commit history & diffs
- Network captures (`.pcap`)
- Images (screenshots of customer environments)
- Spreadsheets, presentations

No single "blocklist.txt": domains are embedded throughout

Methodology: Domain Extraction Pipeline

We built a pipeline to extract domains from all file types in the leak:



Specialized handlers:

- **Images:** Tesseract OCR
- **Archives:** Recursive extraction (RPM, ZIP, TAR, JAR)
- **Git:** Current state + commit history diffs
- **PCAP:** tshark PDML mode

Resolution: DNS resolution using 1.1.1.1 from uncensored network

Measurement Methodology

Vantage Points:

- China (Guangzhou, Nanjing)
- Pakistan
- Myanmar
- Algeria

Unable to access: Kazakhstan, Ethiopia (known Geedge customers)

DNS Injection (China):

- GFW injects fake DNS responses even for non-existent resolvers
- Send query to `1.2.3.4`; any response = censored
- **Highly reliable** indicator

SNI-based TLS (Others):

- Censor terminates TLS handshake if SNI matches blocklist
- Detected via early EOF; **25 tests per domain**

Comparison Lists

Tranco Top 1M (Oct 2025)

- Aggregates: Chrome UX Report, Cloudflare Radar, Cisco Umbrella, Majestic Million
- 30-day rolling window for stability
- Represents *globally popular* domains

Used in censorship research to measure blocking of mainstream content

Citizen Lab Test Lists

- Hundreds of curated URLs per country (28K combined) across 30 categories
- Political criticism, human rights, LGBT, religion, news media, etc.
- Country-specific + global lists
- Curated by volunteers

Used by OONI, Censored Planet, ICLab

Domain Distribution in Tranco Rankings

Tranco Rank	GNL Domains
Top 100	26
Top 1,000	194
Top 10,000	2,520
Top 100,000	38,650
Top 1,000,000	271,735
Not in Tranco	**6,643,531**

Domain Distribution in Tranco Rankings

Tranco Rank	GNL Domains
Top 100	26
Top 1,000	194
Top 10,000	2,520
Top 100,000	38,650
Top 1,000,000	271,735
Not in Tranco	**6,643,531**

96% of GNL domains are outside the Tranco top 1M

GNL Overlap with CitizenLab Test Lists

How much of CitizenLab's curated lists appear in the GNL? *(Note: Ethiopia and Kazakhstan are not measurement vantage points; overlap is GNL-side only)*

CitizenLab List	Total URLs	In GNL	Overlap %
Global	1,678	632	37.7%
China	549	202	36.8%
Ethiopia	191	77	40.3%
Kazakhstan	546	190	34.8%
Pakistan	616	214	34.7%
Myanmar	860	121	14.1%
Combined	**27,646**	**8,068**	**29.2%**

The GNL captures 30-40% of country-specific sensitive domains but also contains many domains *not* in CitizenLab.

Censorship Measurements

Location	CL Local	CL Combined (37,919)	Tranco (1,000,000)	GNL (6,915,266)
China (GZ/NJ)	243/589 (41.3%)	2,696 (7.1%)	7,821 (0.8%)	**218,339 (3.2%)**
Pakistan	28/670 (4.2%)	617 (1.6%)	19,406 (1.9%)	**113,796 (1.6%)**
Myanmar	20/875 (2.3%)	109 (0.3%)	1,713 (0.2%)	**3,131 (0.05%)**
Algeria	22/403 (5.5%)	71 (0.2%)	86 (0.01%)	**299 (0.004%)**

Tranco = popular sites | Citizen Lab = sensitive content | GNL = vendor interest

Key Finding: 298,955 Unique Censored Domains

93.7% of censored GNL domains are NOT in Tranco or Citizen Lab

Key Finding: 298,955 Unique Censored Domains

93.7% of censored GNL domains are NOT in Tranco or Citizen Lab

Location	GNL Censored	Unique to GNL	Unique %
China (GZ/NJ)	218,339	211,746	97.0%
Pakistan	113,796	98,992	87.0%
Myanmar	3,131	2,988	95.4%
Algeria	299	198	66.2%

Country Codes in the GNL

The leak uses internal country/region codes in filenames:

Code	Location	Evidence
E21	Ethiopia	Prior reporting + file context
M22	Myanmar	VPN lists, deployment docs
XJ	Xinjiang	China Unicom (CUCC) SNI data
K23	Kazakhstan?	Deployment references

These codes appear in filenames, folder structures, and internal documentation enabling attribution of specific files to customer deployments.[1]

[1] Country code identification from InterSecLab, "The Internet Coup," 2025. https://interseclab.org/research/the-internet-coup/

File Attribution: Where Do Censored Domains Come From?

Location	Count	File / Description
Multi-country	57,362	`E21-SNI-Top200w.txt`
Multi-country	36,467	`E21-SNI-Top120W-20221020.txt`
Multi-country	24,219	`porn.csv` - Adult content filtering list
Multi-country	13,604	`XJ-CUCC-SNI-Top200w.txt`
China	7,016	`vpn-finder-plugins` - VPN discovery
China	4,810	`Nord VPN server List.txt`
Myanmar	27	`M22-VPN List.html`
Pakistan	68	`Psiphon-CDN_20240430.json`

Key Observations from File Sources

SNI-based Surveillance:

- Largest datasets from MESA lab SNI captures
- 57K+ domains in single monitoring dataset
- Not from popular domain lists
- Appears gathered from network taps

Key Observations from File Sources

SNI-based Surveillance:

- Largest datasets from MESA lab SNI captures
- 57K+ domains in single monitoring dataset
- Not from popular domain lists
- Appears gathered from network taps

VPN Infrastructure Mapping:

- Comprehensive NordVPN server lists
- Psiphon CDN domains
- Country-specific VPN lists (M22, etc.)
- Suggests active circumvention tool tracking

Case Study: Quanzhou Mobile Network Deployment

Found: 白名单网站.txt ("whitelisted websites")

Referenced by deployment documentation showing Geedge software on **mobile telecom network in Quanzhou**:

Allow rules:

- Whitelisted domains

Deny rules:

- Blocked domains
- Fraudulent apps
- User agents (fraud/prostitution)
- Gambling domains
- **APK download interception**

Confirms Geedge software is actively deployed on real telecom infrastructure with both allow and deny rule configurations.

Limitations

- **Rule lists are rare:** Considered sensitive customer data
  - Anything close to a “true blocklist” is mainly from internal discussions of customer environments
- **Overlap with research lists:** GNL contains Alexa, SecRank lists (for MESA research)
  - Both very stale. Alexa is long-discontinued, and SecRank seems to have not updated in a while.
- **PDF processing:** Not yet implemented
- **OCR:** Tesseract is not quite SOTA here

Future Work

- **IP address extraction** from the GNL
- **PDF analysis** and improved OCR
- **Larger domain lists:** Common Crawl, ICANN CZDS[1] may have more overlap
- **Topic analysis:** Categorize unique censored domains

[1] GFWatch uses these

Conclusion

The Geedge Networks leak provides **rare ground truth** on commercial censorship operations.

Conclusion

The Geedge Networks leak provides **rare ground truth** on commercial censorship operations.

298,955 censored domains not in our usual domain lists.

Conclusion

The Geedge Networks leak provides **rare ground truth** on commercial censorship operations.

298,955 censored domains not in our usual domain lists.

Incorporating vendor-leaked data **complements** existing measurement methodologies.

Questions?

Contact:

Jade Sheffey

jsheffey@cs.umass.edu

Code/Data:

Coming soon, feel free to talk to me.

Acknowledgments:

NSF CNS-2333965

DARPA Young Faculty Award