**Analyzing Internet Censorship of Communication Platforms Using OONI Data and Big Data Technologies**

Istanbul Technical University

**ABSTRACT**

This project analyzes internet censorship targeting communication apps like Signal, Facebook Messenger, WhatsApp and Telegram using OONI measurement data. We developed a Spark-based pipeline to process large-scale JSONL files, identifying failure patterns such as DNS errors and TCP resets that indicate blocking. Results are stored in Parquet format and visualized through Elasticsearch. Our findings reveal regional censorship trends and common blocking methods, contributing to a clearer understanding of global digital repression.

**INTRODUCTION**

Internet censorship is a growing concern in the digital age. While open communication is considered a cornerstone of democratic societies, several states implement content control policies under the pretense of national security, social order, or cultural preservation. These restrictions often manifest through technical methods such as IP blocking, DNS tampering, keyword filtering, and throttling access to specific platforms—particularly encrypted messaging services that challenge surveillance capabilities.

To study these practices, organizations like OONI collect and publish network measurement data from around the world. OONI Probe tests assess the accessibility of popular websites and services by performing DNS queries, HTTP requests, TCP connection attempts, and more. The resulting data, stored in JSONL format, reflects both successful and failed connection attempts—offering empirical evidence of potential censorship.

Due to the massive volume and complex nested structure of OONI data, this project leverages Apache Spark for distributed processing and transformation. The cleaned and filtered results are stored in Parquet format for performance and scalability, with selected insights indexed into Elasticsearch for visualization and further analysis. By grouping results by country, blocking method, and date, we aim to provide both quantitative and qualitative insights into the global landscape of digital censorship.

**LITERATURE REVIEW**

Internet censorship has become a pressing research area, particularly as states increasingly interfere with access to communication tools and online content. Measurement platforms like OONI and Censored Planet have emerged to track and analyze censorship globally, each with unique methodologies and strengths.

The Censored Planet platform (Raman et al., 2020) introduced a scalable, remote measurement framework that collects data from over 95,000 vantage points worldwide. Its modular system enables widespread and consistent monitoring, capturing subtle and previously undetected censorship events across different countries. The project emphasizes the importance of scalable, automated monitoring to understand national trends and long-term shifts in censorship practices.

More recently, Raman et al. (2023) highlighted the analytical challenges in interpreting large-scale measurement data. They stressed the need for robust analysis pipelines that can contextualize findings and support policy-oriented insights. This underscores a broader limitation in the field: while data collection is expanding, scalable tools for analysis and visualization remain underdeveloped.

Building on this context, our project leverages OONI’s extensive datasets and applies Apache Spark to enable large-scale, structured analysis of censorship targeting communication apps such as Signal, Facebook Messenger, WhatsApp and Telegram. In doing so, we aim to contribute new insights into when, where, and how such platforms are blocked—framing technical observations within broader questions of internet freedom and digital rights.

**OUR APPROACH**

In this project, we processed and visualized a massive dataset from the Open Observatory of Network Interference (OONI), consisting of approximately 90 TB of semi-structured .jsonl.gz files and around 62 million documents. Initially, we encountered performance issues when attempting to read the entire dataset directly using Apache Spark on AWS EMR, as the job began with a time-consuming file listing operation. To address this, we developed a custom Python script to extract metadata such as file paths and timestamps. During this phase, we also filtered out irrelevant files, reducing the dataset to 26 TB and 14 million files, and stored the metadata in a dedicated S3 location.

Using this metadata, we executed optimized Spark jobs to bulk load only the relevant files, applying a predefined schema to structure the data into DataFrames. From the raw timestamps, we derived and stored additional day, month, and year fields to support filtering and partitioning. Since the dataset included network measurement data from four messaging applications—WhatsApp, Telegram, Facebook Messenger, and Signal—we designed a normalized schema to unify the data structure for consistent analysis.

To address the small files problem, we repartitioned the DataFrame to ensure each output file was approximately 128 MB. The processed data was then saved back to S3 in Snappy-compressed Parquet format, partitioned by date, country, and application to enable efficient querying.

For visualization, we deployed a scalable Elasticsearch cluster on AWS EC2 using Docker, consisting of one master node with Kibana and three data nodes. We then used Spark to batch-load seven years of data (2019–2025) into the cluster.

To ensure the dashboards remain up to date, we automated the data ingestion using Apache Airflow, which runs daily Spark DAGs at 04:00 UTC, shortly after the source data becomes available (typically by 02:00 UTC). This pipeline ensures that Elasticsearch and Kibana visualizations are continuously refreshed with minimal latency.

**FAILURE ANALYSIS**

Some of the technical causes behind test failures are caused by governmental censorship and some indicate network issues or other miscellaneous errors. These failures are consistent with **intentional blocking,** often via DNS poisoning, TCP reset injections, or deliberate timeout strategies by state-level actors.

* dns\_nxdomain\_error

Occurs when DNS queries return a "non-existent domain" response.

* connection\_reset

TCP reset packets are injected to forcibly terminate a connection.

* ssl\_unknown\_authority

Indicates that a forged SSL certificate was presented, possibly by a **man-in-the-middle (MitM)** device.

* generic\_timeout\_error

If consistently observed from multiple vantage points in one country, this may indicate **throttling or silent blocking.**

* connection\_refused

This may be a sign of **filtering via IP blocks or middleboxes.**

**VISUALISATION & INTERPRETATION**

We built an interactive dashboard to visualize trends, geographical distribution, and techniques of internet censorship based on the OONI measurement data. Our goal was to transform raw measurements into an informative tool for public awareness, research, and policy discussions.

1. A graph of a bar graph

   AI-generated content may be incorrect.**Signal**

Figure 1: Signal Censor Graph

Signal shows generally high accessibility across countries between 2021 and 2023. However, there is a significant drop in accessibility starting around late 2023 and early 2024, including a few months with negative values, indicating complete failure or censorship in some regions. This might suggest country-level blocking or technical issues leading to widespread failure.

1. A graph of a bar chart

   AI-generated content may be incorrect.**Whatsapp**

Figure 2: WhatsApp Censor Graph

WhatsApp maintains very high accessibility throughout the entire timeline from 2021 to 2024. There are only minimal fluctuations, with nearly all values close to 1, indicating that most countries had uninterrupted access to the platform.

1. **Telegram**

A graph of a number of years

AI-generated content may be incorrect.

Figure 3: Telegram Censor Graph

Telegram remains largely accessible across most countries, with success-failure ratios close to 1. However, there are a few dips indicating partial accessibility problems, possibly due to local regulations or temporary disruptions in certain regions.

1. A graph of a number of people

   AI-generated content may be incorrect.**Facebook Messenger**

Figure 4: Facebook Messenger Censor Graph

Facebook Messenger was widely accessible between 2021 and 2023. However, from early 2024 onwards, there is a notable and consistent decline in accessibility. The success-failure ratio drops to around 0.5 in some months, suggesting partial censorship or connection failures in several countries.

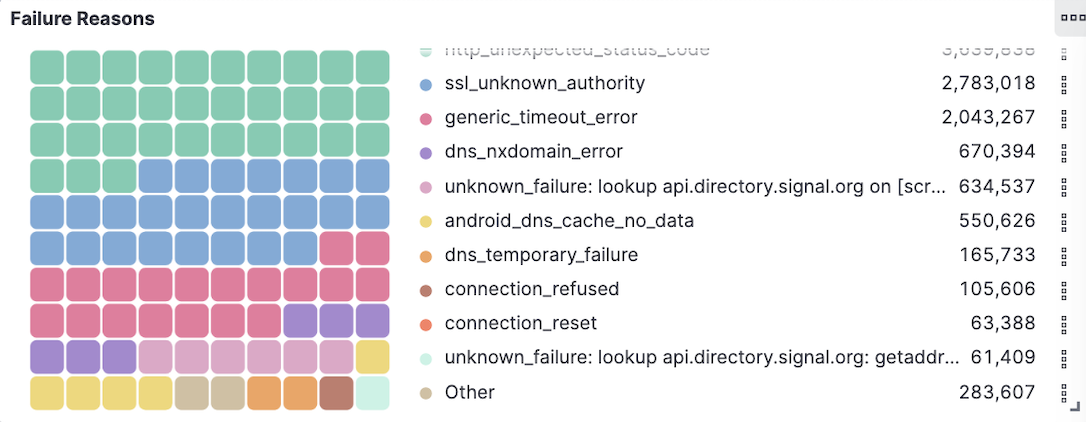
1. **Failure Reasons**

Figure 5:: Failure Reasons

* This grid visualization highlights the technical causes behind test failures.
* Common issues include:
  + ssl\_unknown\_authority: Suggesting interception or invalid SSL certs.
  + dns\_nxdomain\_error: DNS-level censorship where domains are made unreachable.
  + connection\_reset or refused: Typically TCP-level blocking.

1. **TCP vs DNS Blocking Trends Over Time**

A graph of a graph of a number of years

AI-generated content may be incorrect.

Figure 6: TCP & DNS Censor Graph

*TCP Blocking (green line):*

Shows a sharp increase peaking in early 2022, indicating a surge in full connection-level blocking. After that, TCP blocking sharply declined and remained relatively low through 2023 and 2024, suggesting that this method became less commonly used.

*DNS Blocking (blue line):*

Continues at higher and more consistent levels across the entire timeline, peaking around the same time as TCP blocking but maintaining its presence throughout the later years. This suggests DNS manipulation is the more persistent and preferred method of censorship over time.

1. **World Heatmap**

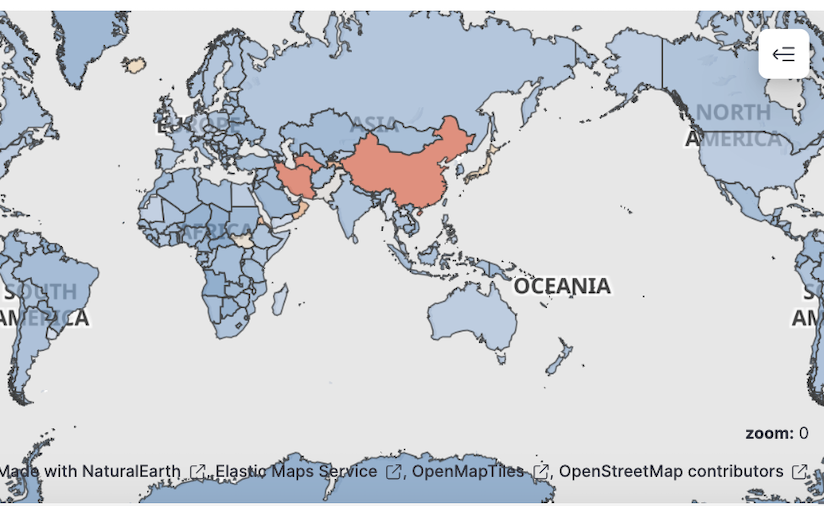
****

Figure 7: Censor Heatmap

This map highlights the geographic scope of internet censorship.

Countries in red have had a high frequency of confirmed censorship events.

Allows a quick glance at global hotspots of repression in the digital domain.

1. **Distribution of Access Attempt Counts**

A pie chart with text

AI-generated content may be incorrect.

Figure 8: Distribution of tests

The attempts are relatively evenly distributed across all four applications, demonstrating their global relevance. This balanced testing coverage ensures that the analysis of accessibility and censorship patterns is fair and comparative.

1. A screen shot of a computer

   AI-generated content may be incorrect.**Accessibility Rate by Turkish Telecommunication Providers**

Figure 9: Turkish Telecommunication Providers

The accessibility of communication platforms in Turkey varies significantly by provider. While some ISPs enable broader access, others may enforce restrictions more frequently, reflecting differences in policy, technical architecture, or regulatory compliance.

**EXPERIMENTS**

These calculations are based on processing 3 months of data for each application.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Setup No** | **Primary Node** | **Core Node(s)** | **Total**  **vCPU** | **Total**  **RAM** | **Estimated Duration (min)** |
| 1 | m5.xlarge (1 unit) | m5.xlarge x 1 | 8 | 32 GiB | 120 |
| 2 | m5.xlarge (1 unit) | m5.xlarge x 3 | 16 | 64 GiB | 45 |
| 3 | m5.xlarge (1 unit) | m5.xlarge x 6 | 28 | 112 GiB | 19 |
| 4 | m5.xlarge (1 unit) | m5.xlarge x 10 | 44 | 176 GiB | 11 |

Table 1: Cluster Configurations and Performance Comparison for 3-Month Data Processing

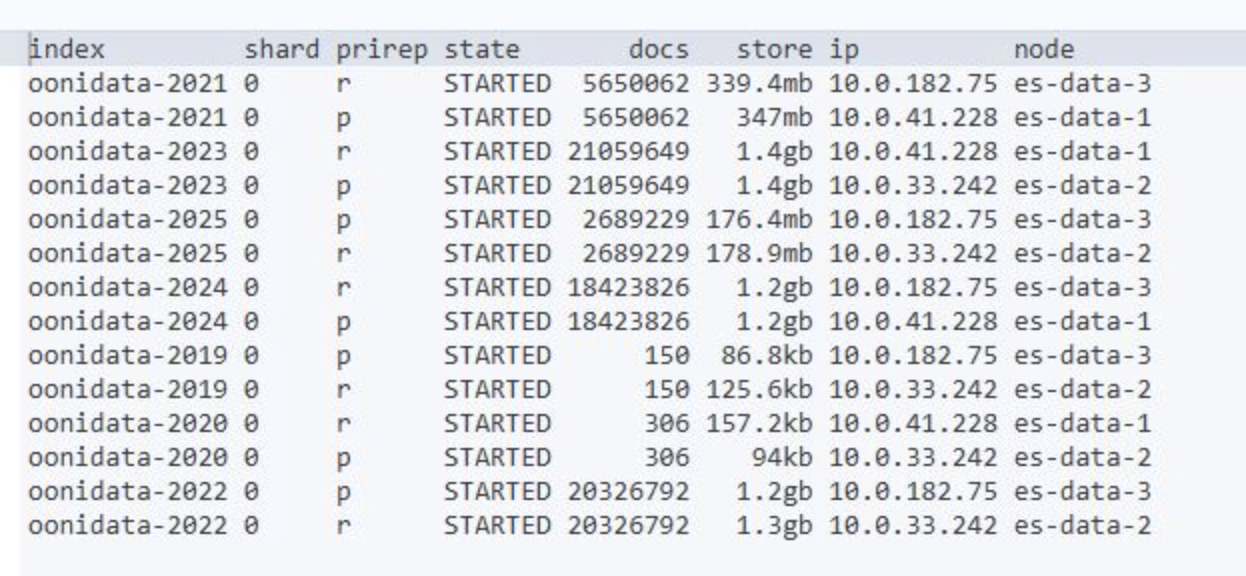


Figure 9: Elasticsearch Shard Allocation and Storage Overview

## Medallion Architecture

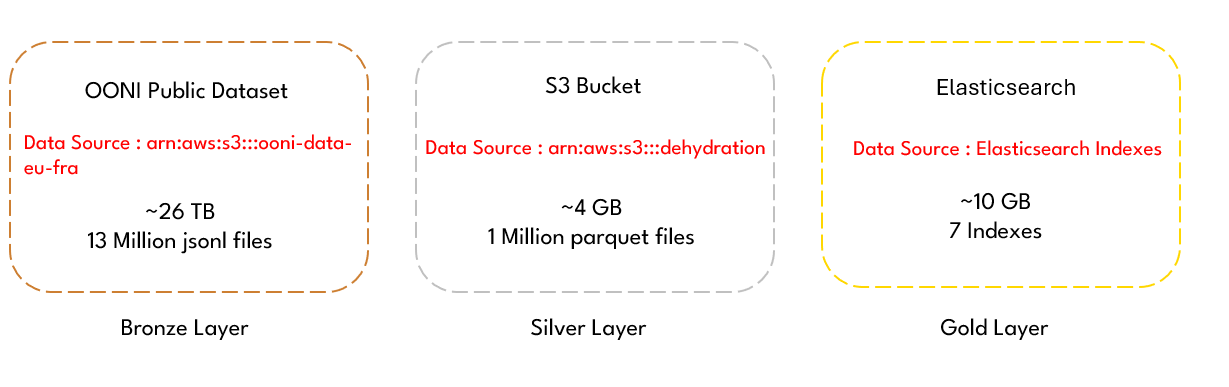


Figure 10 : Medallion Architecture

REFERENCES

1. **Raman et al., 2020 (Censored Planet)**

S. Raman, N. Feamster, D. Choffnes, and M. Schapira, “Censored Planet: An Internet-Wide, Longitudinal Censorship Observatory,” in Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS), 2020, pp. 1–18.

1. **Raman et al., 2023 (Analysis Pipeline)**

S. Raman, N. Feamster, D. Choffnes, and M. Schapira, “The Censorship Analyzer: Designing and Deploying an Analysis Pipeline for Censorship Measurement Platforms,” in Proceedings of the USENIX Symposium on Networked Systems Design and Implementation (NSDI), 2023, pp. 1–15.