Mapping legally withheld web content: Tooling for HTTP 451

Alp Toker @atoker alp@netblocks.org 28 June 2017
NetBlocks.org Internet Health Working Group

Background

Based on lessons learnt from three iterations of a prototype tool for detection of legally withheld / geoblocked world wide web resources, this document outlines our conclusions and recommendations for a quality implementation crawling tool based on emerging standards and industry practices. This document does not endorse or recommend the use of HTTP 451 code in deployment.

Purpose

The purpose of this tool is to efficiently scan for web content that is identified blocked for legal reasons.

Scope

- Detect HTTP error codes, primarily HTTP Code 451
- Use non-blocking IO
- Scale from low-end embedded ARM (e.g. Raspberry Pi) up to server hardware with a single code base
- Lightweight textual JSON-based output format for downstream processing and aggregation
- Deterministic behaviour and locked dependencies
- Semantic versioning and dependency locking for reproducible builds and deterministic runs

Requirements and considerations

- Requirements:
  - The tool should support throttling, rate limiting and customised crawling logic for different sites based on well-known search patterns.
  - The tool should allow integration with forensic packet capture infrastructure (e.g. with NetBlocks PCAP framework and UNIX-like TCP/IP stacks, this is achieved by binding the client socket to a specific port prior to connecting)
  - Throttling, rate limiting and crawling logic ideally be built upon an existing quality implementations
- Nice-to-have and future extensibility:
  - Custom DNS resolver to circumvent OS-defined values which may be user-overridden
  - Future: Detect and classify other forms of low-level network interference (may entail use of low-level or platform-specific TCP stack facilities)
  - Future: Fine-grained per-request access-timings
  - Future: TLS certificate collection and analysis

Strategy Definition Formats

By (1) limiting the search space to URL patterns that are known to be typically blocked by a provider or intermediary, and (2) guiding the crawler towards likely blocked content we can increase the effectiveness of the tool and reduce network traffic.

Crawling strategies should be defined per-site with a simple and stable format. A “strategy” is a well-known crawling pattern that guides (1) the scope of the crawl (2) the order of a crawl.

Example:

A strategy for crawling geoblocked reddit forums (subreddits) may be defined with:

- **Limit:** A regular expression that limits requests to the regular expression matching toplevel subreddit URLs: ^/r/[^/]*$
- **Ordering:** First scan for known-blocked content, then e.g. using readily available NSFW subreddit list.