WEB CRAWL REFUSALS: INSIGHTS FROM COMMON CRAWL

23 JULY 2025, MAPRG SESSION ON AI CRAWLER TRAFFIC IMPACTS

MOSTAFA ANSAR
RADBOUD UNIVERSITY

ANNA SPEROTTO
UNIVERSITY OF TWENTE

RALPH HOLZ
UNIVERSITY OF MÜNSTER

Radboud University

UNIVERSITY OF TWENTE.

Universität Münster

INTRODUCTION

- Automated traffic now makes up 30–50% of all web traffic[1,2]
- AI crawlers are a major driver behind the recent surge
- Previously, crawlers indexed content and referred users: a fair exchange
- Now, their scraping imposes heavy infrastructure load, but offers little in return
  - More on this in later talks
- Bots now seen as exploitative and unwelcome
- Our focus: what does this mean for research that relies on web crawling?

[1] https://radar.cloudflare.com/bots

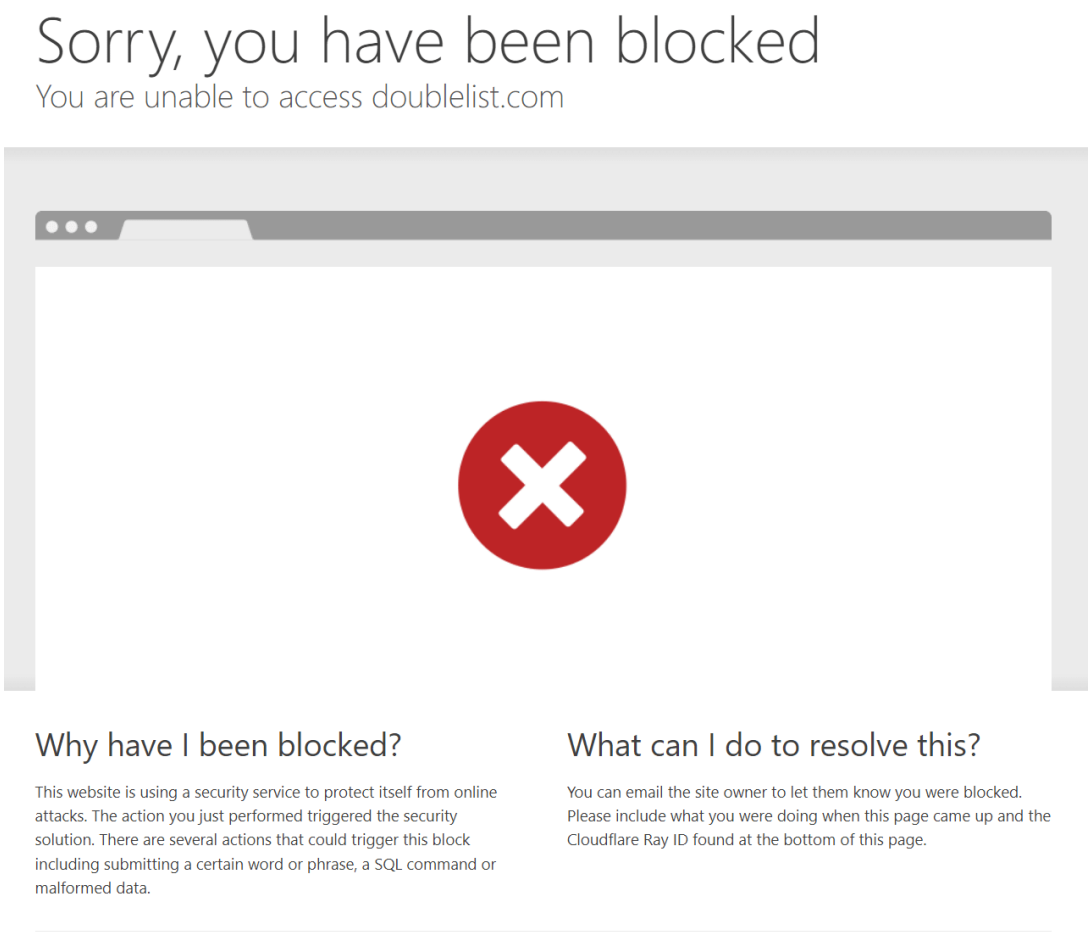
[2] https://www.imperva.com/resources/resource-library/reports/2025-bad-bot-report

INTRODUCTION

- As bots are becoming unwelcome, blocking bots becomes more widespread
- Crawling the web to collect research data is becoming increasingly difficult
- This undermines tools like Common Crawl
- Its data used in various research fields (cited 10,000+ times[1]), not just AI training
- Common Crawl not exactly an AI crawler, but sometimes labelled as an AI user agent[2]
- In late 2023, we conducted research on how blocks (or “refusals”) affect Common Crawl
  - Before the recent uptake in AI crawler traffic
  - CommonCrawl’s coverage was already suffering from server-side blocks

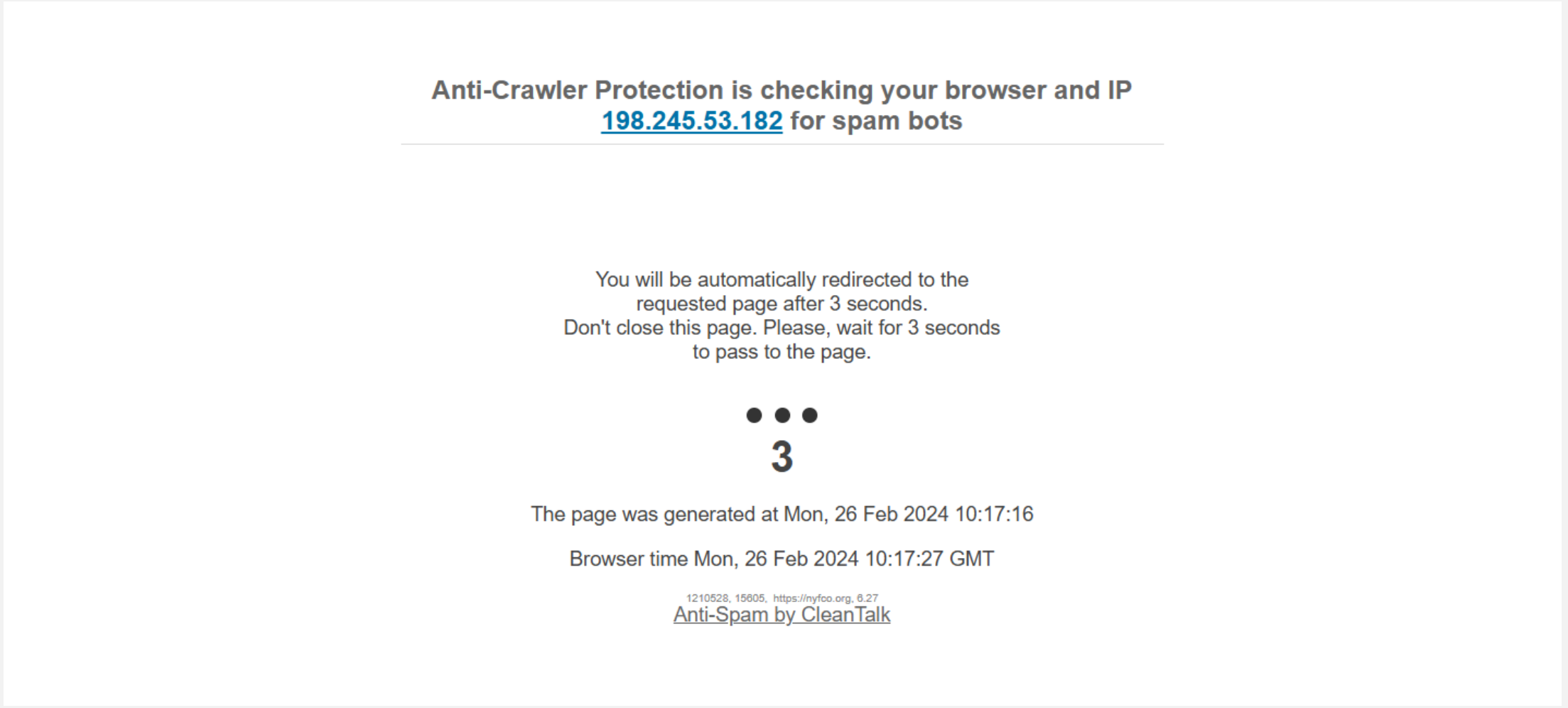
[1] https://commoncrawl.org
[2] https://radar.cloudflare.com/ai-insights

WHAT ARE REFUSALS



https://www.reddit.com/r/CloudFlare/comments/13zdbsr/why_does_cloudflare_block_me/

WHAT ARE REFUSALS



https://archive.ph/G8PXG

WHAT ARE REFUSALS

https://docs.crowdsec.net/u/bouncers/wordpress/

WHAT ARE REFUSALS

RESEARCH QUESTIONS & CONTRIBUTIONS

- Refusals → potentially skews research data

- **How often and why does CommonCrawl face refusals?**

- **Can crawlers like CommonCrawl identify *refusals* to react accordingly?**
  - Different shapes and forms
  - Hard to distinguish between general errors and *refusals*

- **Our contribution:**
  - Analysis of explicit refusals in a Common Crawl snapshot
  - Evidence of diverse refusal signalling by websites
  - Recommendations for crawler adaptability

[1] https://commoncrawl.org/

METHODOLOGY

- **Scope:** Websites that did not disallow “CCBot” in their robots.txt
  - Since CommonCrawl respects robots.txt
- **Dataset:** CC-MAIN-2023-50 compiled in November and December 2023
- **Overview of steps:**
  1. Parse non-200 responses
  2. Extract textual contents
  3. Manually create regular expressions for capturing refusals
  4. Capture and categorize refusals
  5. Analyze refusal transience

METHODOLOGY

- **Regular Expressions:**
  - Built iteratively to capture *explicit* refusals (not general error pages)
  - Initial Keywords ⟶ Loose Regular Expressions ⟶ Fine-grained Regular Expressions
  - Final Set: **147 Regular Expressions**
  - Manually labelled based on the content it captures:
    - *Type:* Block, Challenge, Checking, Require_JS, None
    - *Reason:* Security/Malicious, Excessive/Suspicious, etc.
    - ***Tag:*** A platform, tool or security provider, e.g. Cloudflare, Shopify, ModSecurity

METHODOLOGY

- **Example:**
  - **cloudflare please enable cookies. sorry, you have been blocked you are unable to access artbattle.com why have i been blocked? this website is using a security service to protect itself from online attacks. the action you just performed triggered the security solution. there are several actions that could trigger this block including submitting a certain word or phrase, a sql command or malformed data. what can i do to resolve this? you can email the site owner to let them know you were blocked. please include what you were doing when this page came up and the cloudflare ray id found at the bottom of this page. cloudflare ray id: 830aa4ef09258f0a • your ip: click to reveal 3.238.180.174 • performance & security by cloudflare**

RegEx: ?

Type: ?

Reason: ?

Tag: ?

METHODOLOGY

- **Example:**
  - cloudflare please enable cookies. sorry, you have been blocked you are unable to access artbattle.com why have i been blocked? this website is using a security service to protect itself from online attacks. the action you just performed triggered the security solution. there are several actions that could trigger this block including submitting a certain word or phrase, a sql command or malformed data. what can i do to resolve this? you can email the site owner to let them know you were blocked. please include what you were doing when this page came up and the cloudflare ray id found at the bottom of this page. cloudflare ray id: 830aa4ef09258f0a • your ip: click to reveal 3.238.180.174 • performance & security by cloudflare

RegEx: `^.{20,250}us(ing|es) a security service (to|for) protect(ion)? (itself from|against) online attacks`

Type: ?

Reason: ?

Tag: ?

METHODOLOGY

- **Example:**
  - cloudflare please enable cookies. sorry, you have been blocked you are unable to access artbattle.com why have i been blocked? this website is using a security service to protect itself from online attacks. the action you just performed triggered the security solution. there are several actions that could trigger this block including submitting a certain word or phrase, a sql command or malformed data. what can i do to resolve this? you can email the site owner to let them know you were blocked. please include what you were doing when this page came up and the cloudflare ray id found at the bottom of this page. cloudflare ray id: 830aa4ef09258f0a • your ip: click to reveal 3.238.180.174 • performance & security by cloudflare

RegEx: `^.{20,250}us(ing|es) a security service (to|for) protect(ion)? (itself from|against) online attacks`

Type: **block**

Reason: ?

Tag: ?

METHODOLOGY

- **Example:**
  - cloudflare please enable cookies. sorry, you have been blocked you are unable to access artbattle.com why have i been blocked? this website is using a security service to protect itself from online attacks. the action you just performed triggered the security solution. there are several actions that could trigger this block including submitting a certain word or phrase, a sql command or malformed data. what can i do to resolve this? you can email the site owner to let them know you were blocked. please include what you were doing when this page came up and the cloudflare ray id found at the bottom of this page. cloudflare ray id: 830aa4ef09258f0a • your ip: click to reveal 3.238.180.174 • performance & security by cloudflare

RegEx: `^.{20,250}us(ing|es) a security service (to|for) protect(ion)? (itself from|against) online attacks`

Type: **block**

Reason: **security/malicious**

Tag: ?

METHODOLOGY

- **Example:**
  - cloudflare please enable cookies. sorry, you have been blocked you are unable to access artbattle.com why have i been blocked? this website is using a security service to protect itself from online attacks. the action you just performed triggered the security solution. there are several actions that could trigger this block including submitting a certain word or phrase, a sql command or malformed data. what can i do to resolve this? you can email the site owner to let them know you were blocked. please include what you were doing when this page came up and the cloudflare ray id found at the bottom of this page. cloudflare ray id: 830aa4ef09258f0a • your ip: click to reveal 3.238.180.174 • performance & security by cloudflare

RegEx: `^.{20,250}us(ing|es) a security service (to|for) protect(ion)? (itself from|against) online attacks`

Type: **block**

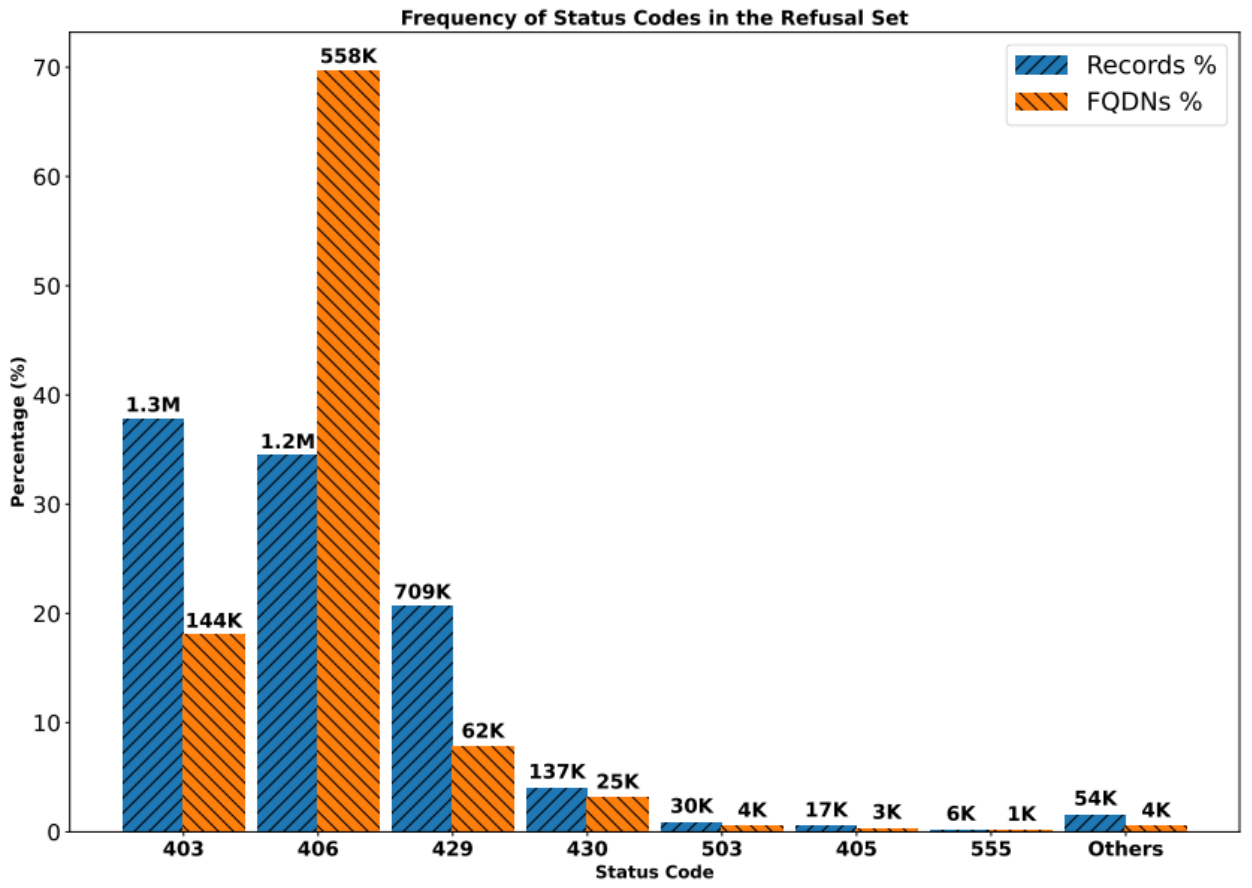
Reason: **security/malicious**

Tag: **cloudflare**

KEY FINDINGS

- ~**800,000** sites had explicit refusals
  1.68% of all sites in the snapshot

- **31** different status codes
- **Top ones:**
  - **403 (Forbidden)**
  - **406 (Not Acceptable**)
  - **429 (Too Many Requests)**
  - **430 (Unassigned)**

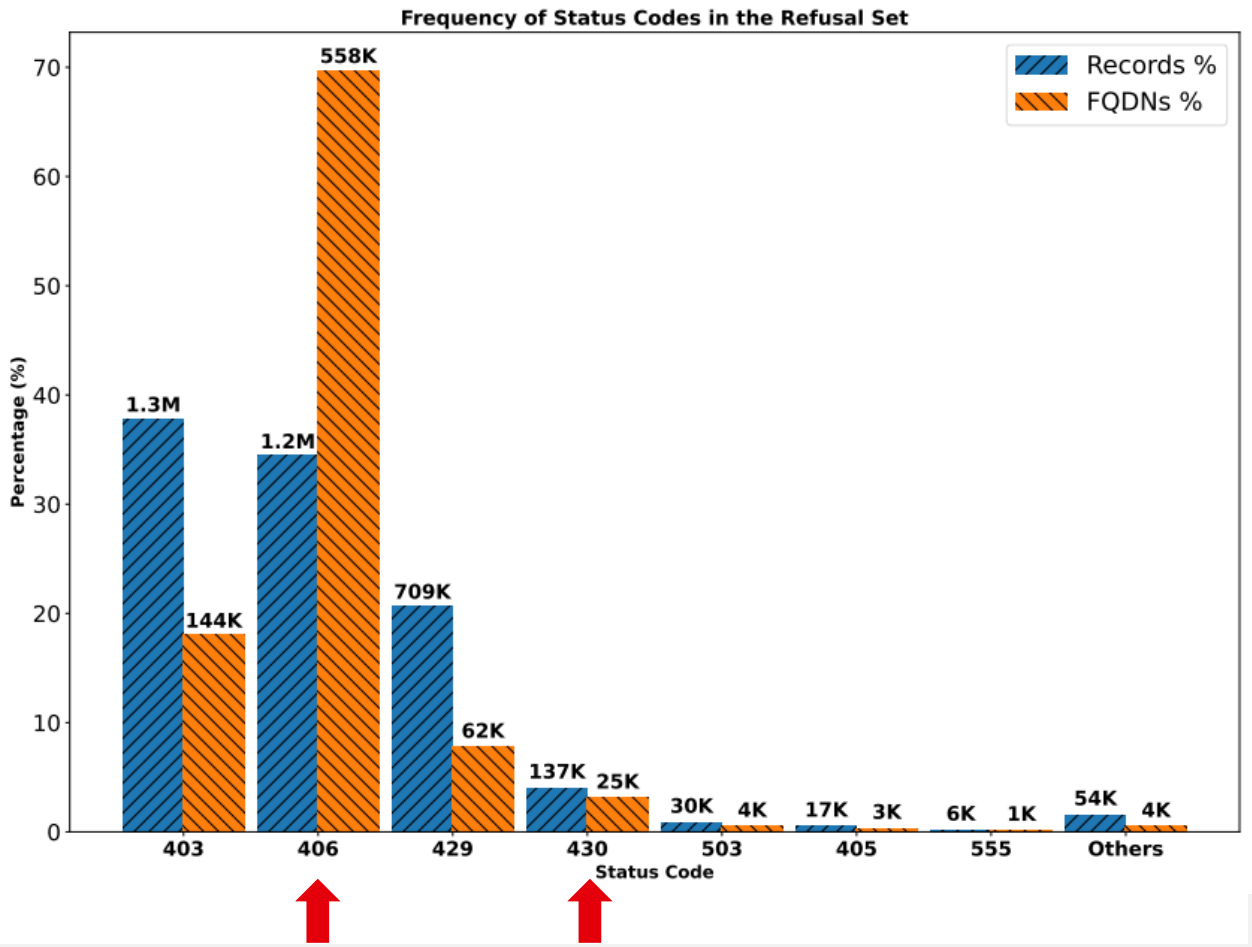


KEY FINDINGS

- ~**800,000** sites had explicit refusals
  1.68% of all sites in the snapshot

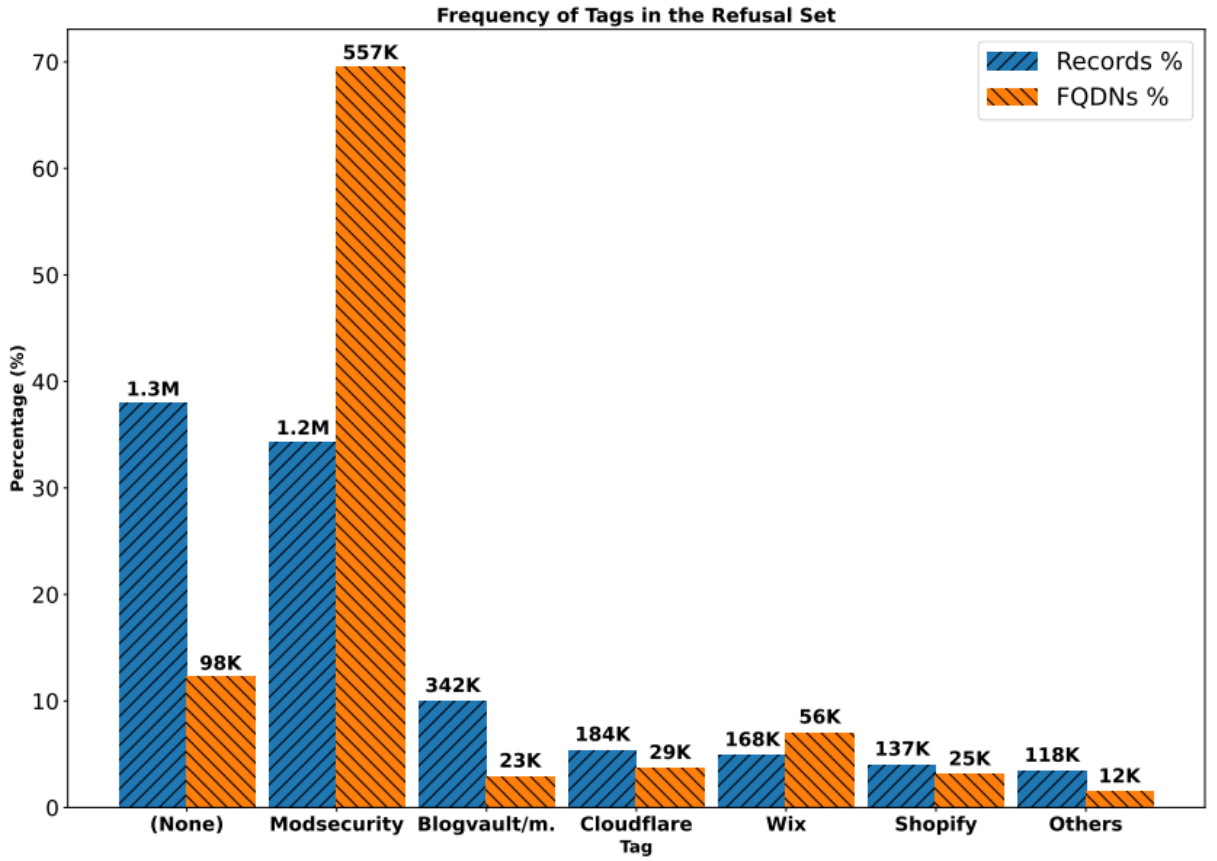
- **31** different status codes
- **Top ones:**
  - **403 (Forbidden)**
  - **406 (Not Acceptable**)
  - **429 (Too Many Requests)**
  - **430 (Unassigned)**

- Unexpected:
  - **406**: For representation errors
  - **430**: Used by Shopify



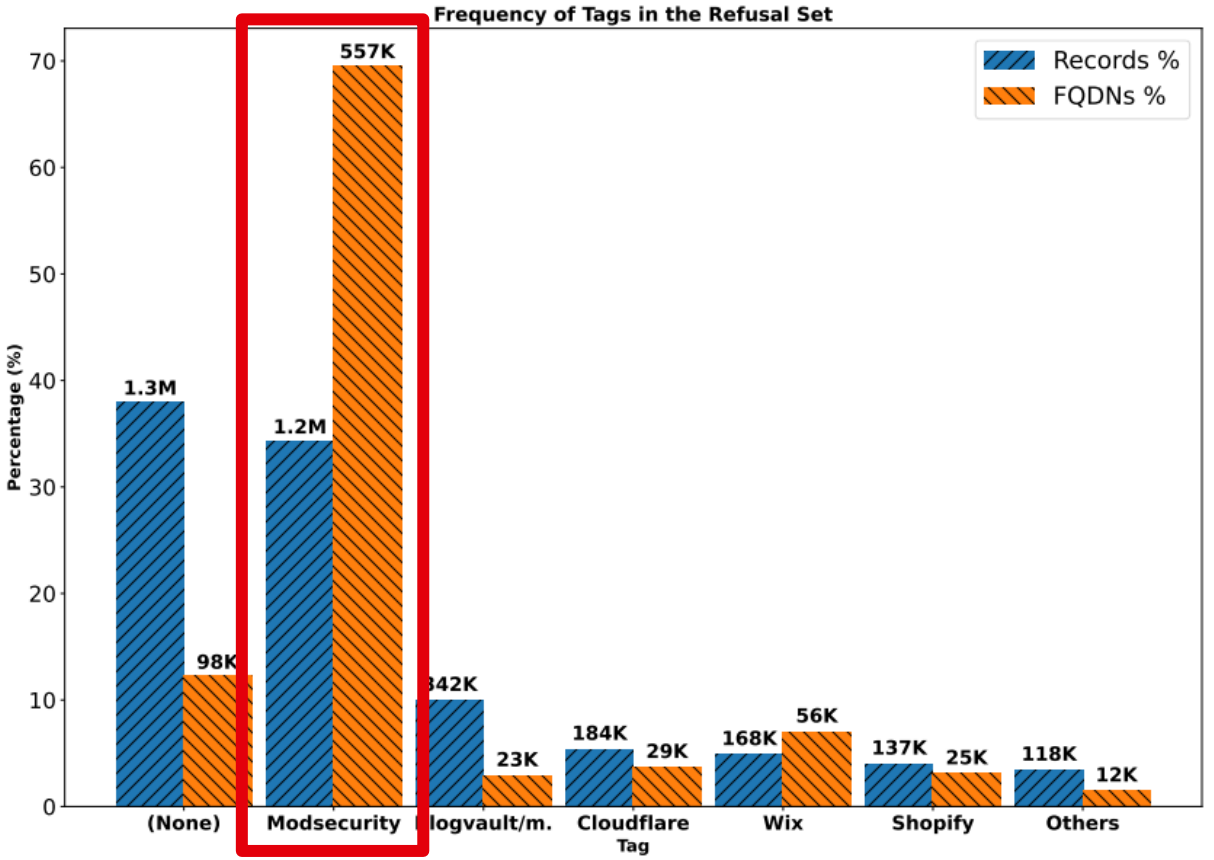
KEY FINDINGS

- **28** different providers/platforms
- **Top ones:**
  - ModSecurity
  - Blogvault/Malcare (Wordpress plugin)
  - Cloudflare



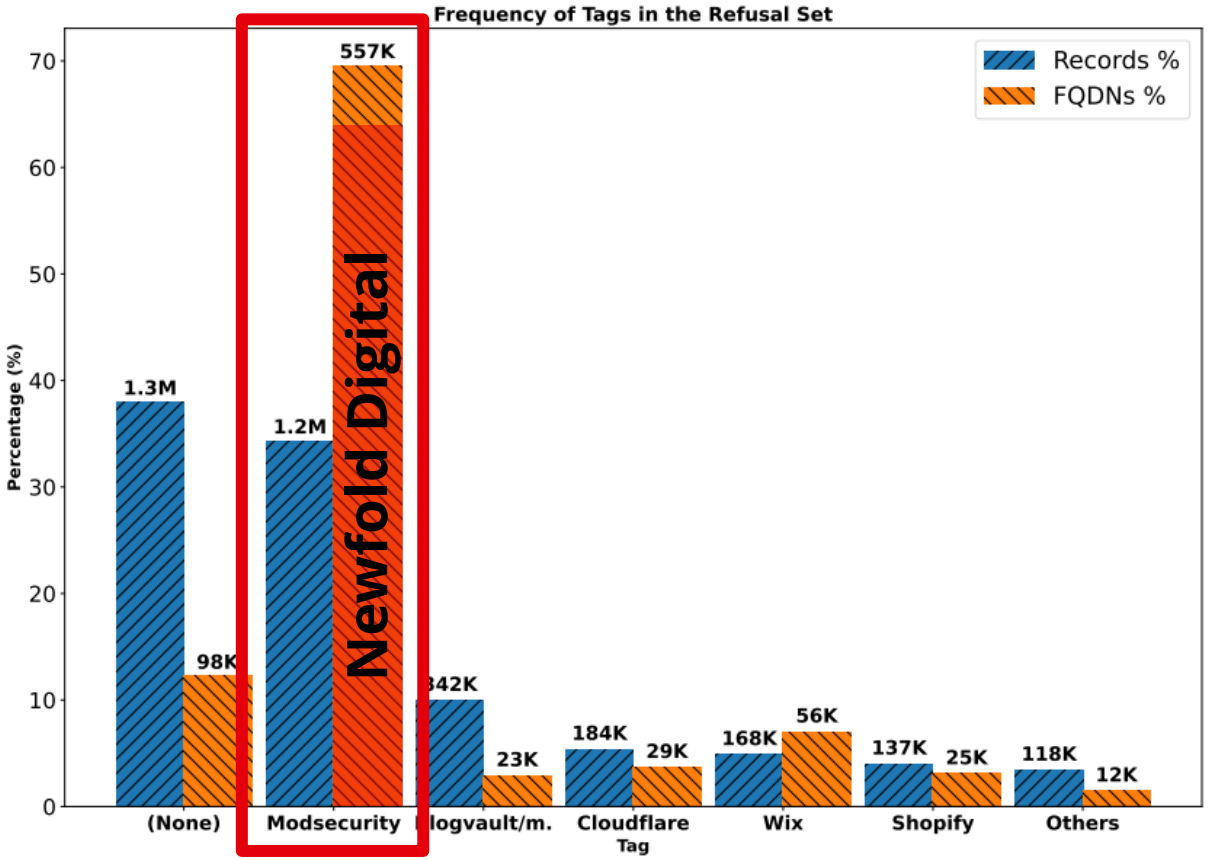
KEY FINDINGS

- **28** different providers/platforms
- **Top ones:**
  - ModSecurity
  - Blogvault/Malcare (Wordpress plugin)
  - Cloudflare
- **ModSecurity**
  - **70%** of all refusing domains



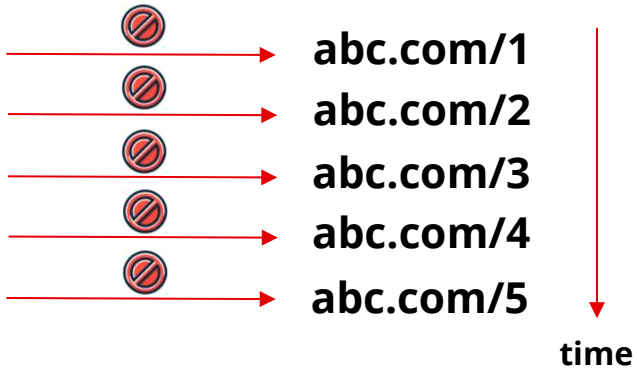
KEY FINDINGS

- **28** different providers/platforms
- **Top ones:**
  - ModSecurity
  - Blogvault/Malcare (Wordpress plugin)
  - Cloudflare
- **ModSecurity**
  - **70%** of all refusing domains
  - **92%** of them on a single hosting conglomerate (Newfold Digital)
  - They block “CCBot”
  - Centralized aggressive policy



KEY FINDINGS

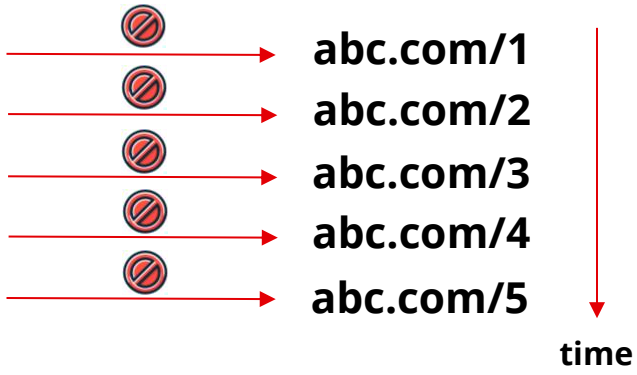
- **Transience**
  - **~80%** of blocking websites block on every request
  - Unexpected:
    - **429 (Too Many Requests)** meant for rate-limiting but used for non-transient refusals by **40%** of websites[1]



[1] Those FQDNs whose refusals are communicated via status code 429

KEY FINDINGS

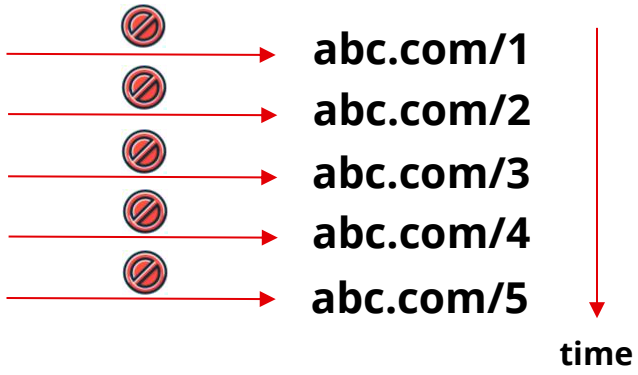
- **Transience**
  - **~80%** of blocking websites block on every request
  - Unexpected:
    - **429 (Too Many Requests)** meant for rate-limiting but used for non-transient refusals by **40%** of websites[1]
  - Some status codes: more transient
    - Non-transient: e.g. **406**, **405**
    - Transient: e.g. **430**, **403**

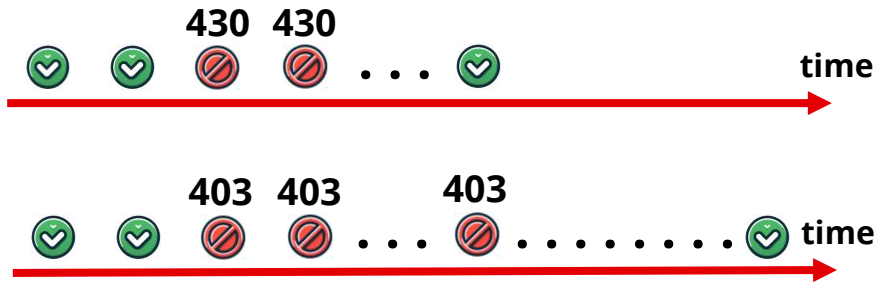


[1] Those FQDNs whose refusals are communicated via status code 429

KEY FINDINGS

- **Transience**
  - **~80%** of blocking websites block on every request
  - Unexpected:
    - **429 (Too Many Requests)** meant for rate-limiting but used for non-transient refusals by **40%** of websites[1]
  - Some status codes: more transient
    - Non-transient: e.g. **406**, **405**
    - Transient: e.g. **430**, **403**
  - Some transient ones: shorter time-to-unblock:
    - **430** quickest to resolve (Shopify)
    - **403** slower to resolve





[1] Those FQDNs whose refusals are communicated via status code 429

TAKEAWAYS

- **Refusals Impact Coverage**
  - Not widespread, still a non-negligible challenge to Web crawling coverage
- **Centralization Can Exacerbate the Problem**
  - Newfold Digital: a single rule affects over half a million sites
  - One entity's decision create a massive blind spot for crawlers
- **Impact of AI crawlers**
  - AI crawlers likely to worsen this
    - Cloudflare introduced features for blocking AI bots not long after this research
    - CCBot already considered an "AI Bot" [1]

[1] https://developers.cloudflare.com/bots/concepts/bot/#ai-bots

RECOMMENDATIONS

- **Clearer Standards**
  - HTTP status codes misused → crawlers cannot reliably interpret server “signals”
  - Potential need for a new standard for clear and streamlined signalling to bots
- **Crawling Adjustments**
  - Flexible back-off strategies on HTTP status codes, for instance:
    - Stop if **406 (Not Acceptable)**
    - Wait longer after **403 (Forbidden)**
    - Wait shorter after **430 (Unassigned)**

- **Thank you for your attention.**
- **Contact: mostafa.ansar@ru.nl**
- **Q&A**