Knight–Georgetown Institute

The Case for a Neutral Web Index

Alissa Cooper

The crawling economy is extremely inefficient.

Ex: Duke University study

Data subset	Unique IP addresses	Unique user agents	Avg. bytes scraped per session	Unique ASNs	Total bytes scraped	Total page visits	Unique page visits
All data	231,859	19,250	82,306	8,841	62,713,813,720	761,956	31,665
Known bots	11,291	405	52,612	179	16,706,054,178	317,532	6,347

Bots on university websites in a 6-week period

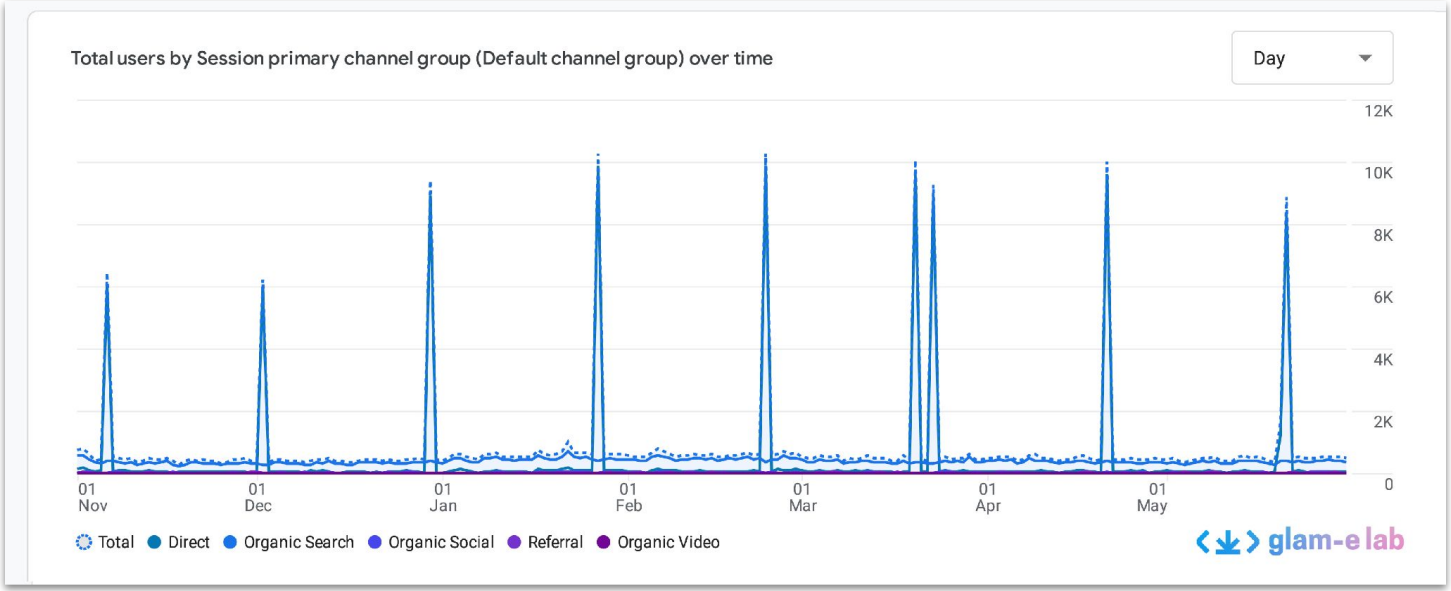
Kim et al. Scrapers Selectively Respect robots.txt Directives: Evidence From a Large-Scale Empirical Study. In *Proceedings of the 2025 ACM Internet Measurement Conference (IMC '25)*, October 28–31, 2025. https://doi.org/10.1145/3730567.3764471

65% of our most expensive traffic comes from bots

Mueller et al. How crawlers impact the operations of the Wikimedia projects. April 1, 2025.


Kyle Wiens
@kwiens
IFIXIT
Hey @AnthropicAI: I get you're hungry for data. Claude is really smart!
But do you really need to hit our servers a million times in 24 hours?
You're not only taking our content without paying, you're tying up our devops resources. Not cool.
11:08 AM · Jul 24, 2024 · 1.6M Views

The information commons is suffering in unique ways.



Weinberg. Are AI Bots Knocking Cultural Heritage Offline? GLAM-E Lab. June 2025.

→ A lot of crawling is duplicative and unnecessary.

What if we had a fairly up-to-date and comprehensive index of the web?

The Neutral Web Index

Neutral Web Index

Web Publisher Registration

Web publishers register to have their content indexed, and on what terms.

Central Clearinghouse

Neutral index becomes the clearinghouse for most bot access.

Controlled Access

Publishers choose to limit direct fetching to human-like accesses (or not).

Neutral web index could become the focal point for …

- Bot authentication
- Access control
- Dealing with the cat-and-mouse game of masquerading bots
- Consuming publisher signals about how and when they want to be re-crawled (a la IndexNow)
- Automated negotiation of payments and licenses
- Value-added features (in the spirit of, e.g., Wikimedia Enterprise)

Web for …

Human consumption

Curated indexing

Bots dedicated to mimicking human inefficiency

Neutral index for …

All other bot consumption

But what about …

… centralization,

… concentration of power,

… monopoly?

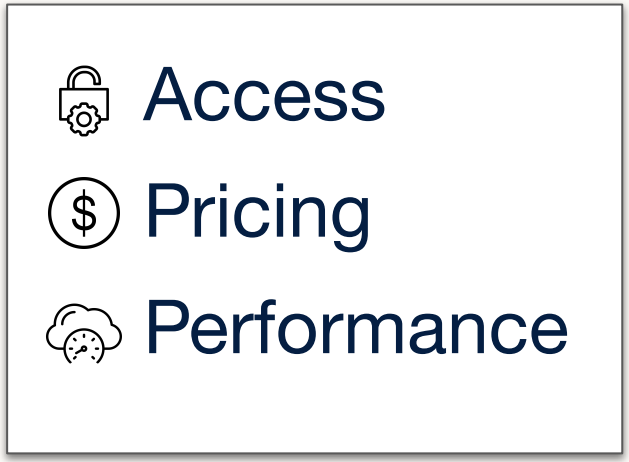
> After having carefully considered and weighed the witness testimony and evidence, the court reaches the following conclusion: **Google is a monopolist**, and it has acted as one to maintain its monopoly.

U.S. District Court Judge Amit P. Mehta
U.S. v. Google Liability Opinion
August 5, 2024

Global snapshot

- **European Commission**: designated Alphabet as the sole “gatekeeper” in online search, Sept 5, 2023
- **UK Competition authority**: “Having carried out an investigation and consulted Google and other stakeholders, we have decided to designate Google as having [strategic market status] in the provision of general search and search advertising,” October 10, 2025
- **European Commission:** opened investigation into possible anticompetitive conduct by Google in the use of online content for AI purposes, Dec 8, 2025
- **Brazilian competition tribunal**: “there is a need to initiate an Administrative Proceeding to investigate possible exploitative abuse of a dominant position,” April 23, 2026

Neutral web index needs some regulatory oversight



> Within thirty (30) days of a Qualified Competitor's certification ..., Google shall make available, at marginal cost, to Qualified Competitors the following data related to Google's Web Search Index:
>
> 1. for each document in the Google Web Search Index, a unique identifier (DocID) …
> 2. a DocID to URL map; and
> 3. for each DocID, the (A) time that the URL was first seen, (B) time that the URL was last crawled, (C) spam score, and (D) device-type flag.

U.S. District Court Judge Amit P. Mehta

U.S. v. Google Final Judgment

December 5, 2025

Can we use competition remedies (or some other mechanism) to reduce inefficiency in the crawling economy?

Thank *You*