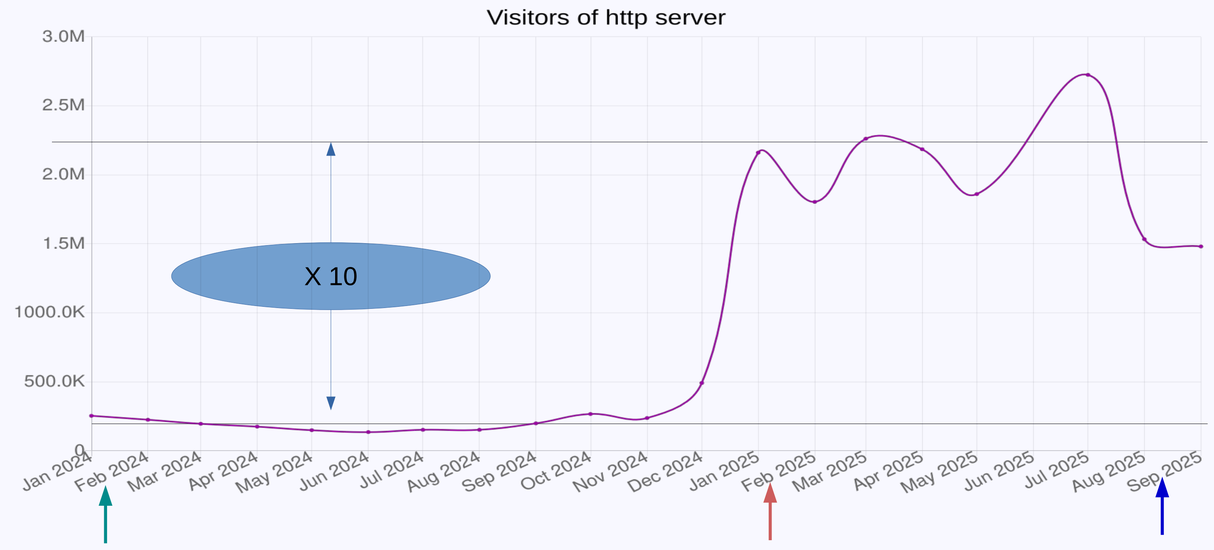
ABUSE OF SERVICE CRAWLING

The SIMBAD astronomical database, has experienced a significant surge in API requests, particularly from automated systems and suspected AI model training pipelines.

OBERTO Anaïs, LANDAIS Gilles

Many IPs origins

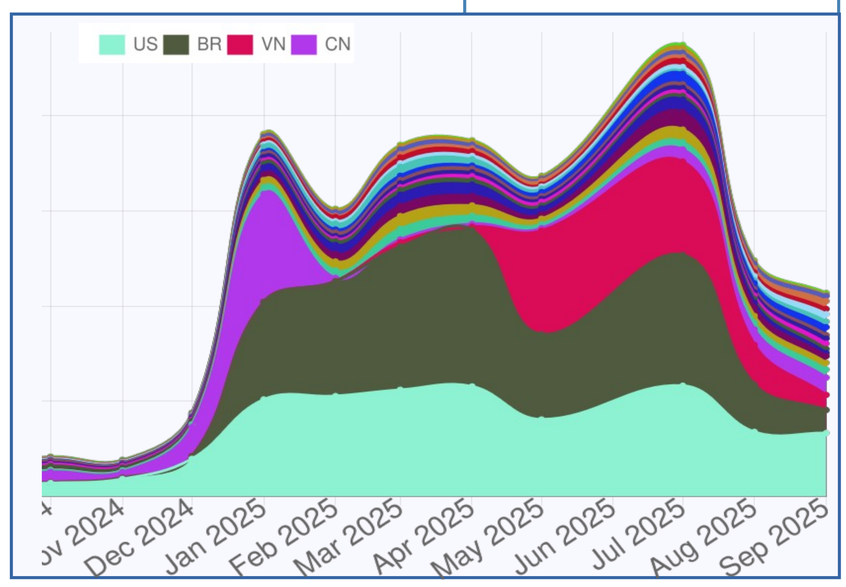


Usually 1month (2024-02)	~ 225K IPs	~ 7 500 IPs / day
Compared to 2025-02	~ 1.8M IPs	~ 60 000 IPs / day
Compared to 2025-09	~ 1.5M IPs	~ 50 000 IPs / day

Max in 2025 : 200 000 IPs /day

Monthly statistics (identified bots removed) reveal a huge number of different IP address accessing the service per month or per day. In mean, it has increase by a factor of 10 in few weeks.

Many countries



The distribution by country of IP origin shows that the increase is strongly linked to the country but changes over time.

The increase is also big (x5) if we disregard the final IP number.

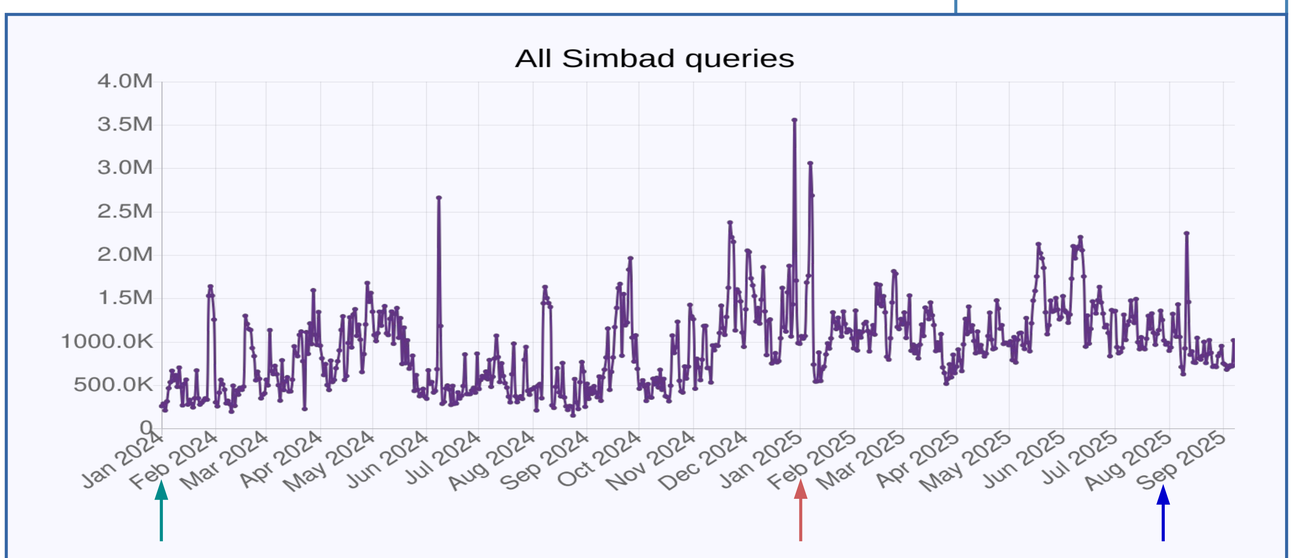
No user-agent

95% of them do not have a web “referer”, and the user-agent description is either empty or “Mozilla xxx-something-xxx”. The user-agent identified as bots “OpenAI / Anthropic / perplexity / FacebookBot / Mistral…” are only about 6% of all queries.

VizieR is experiencing similar behaviour, with a high number of IP addresses (+1489% in 2025).

In addition to traffic generated by bots or AI, the service may be subject to intensive activities from requests made by real users (e.g. >50 requests/second).

Many queries



Usually (2024-02)	~ 0.5M	(0.2M to 1.3M) / day
Compared to 2025-02	~ 1.1M	(0.7M to 1.7M) / day
Compared to 2025-09	~ 0.7M	(0.5M to 1.4M) / day

Max in 2025 : 3.1M / day

We saw a growth of about twice number of queries reaching higher maxima (identified bots removed). The queries do not have a specific signature and are correct queries.

Reactions

These abusive users (real or bots) require regular filtering maintenance to avoid penalising all users.

Simbad and Vizier have to use strategies, including blacklists, HTTP tuning (apache) and application filtering based on request parameters.

Distinguishing between malicious scrapers and legitimate AI agents is difficult, as it is important to prevent abuse while still allowing useful features such as search, citations and assistants.