Good Bot, Bad Bot: Characterizing Automated Browsing Activity

Xigao Li, Babak Amin Azad, Amir Rahmati, Nick Nikiforakis
Stony Brook University

Stony Brook University
Computer Science

What do bots do?

- Web bots are programs that perform web requests, and interact with websites on the Internet.
- While benign bots provide indexing services, content previews or are used for research, attackers use malicious bots to discover vulnerable websites, compromise their servers, and exfiltrate sensitive user data.
- Bots are using evasion techniques such as spoofing User Agents, using automated browsers, or hiding behind proxies to evade bot detection.
- Creating a corpus of bot activity (e.g. to be used for training automated detection systems) is not trivial and has historically been done manually.

Takeaways

> Even unpopular websites receive at least 1,200 requests/day, <2% are benign

> Bots are highly selective, targeting easy-to-exploit endpoints

> 97% bots are built on rudimentary HTTP libraries (e.g. curl), but they pretend to be browsers

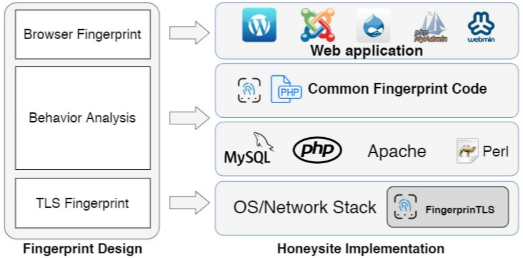
> Only 13% of bot IPs appeared in IP blocklists

> TLS fingerprinting is effective against cloaking and evasions

> Exploits that go public are quickly abused - Even if you are hosting an unpopular website, deciding to patch a vulnerability over the weekend may already be too late.

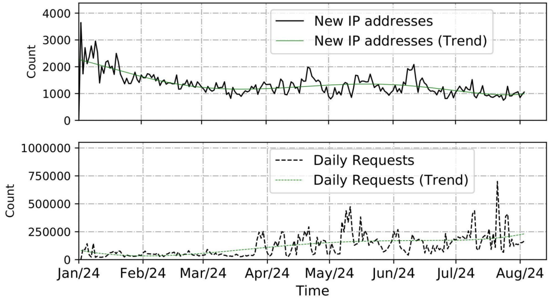
How can we build a bot-only dataset?

- We design and build Aristaeus, a system that provide flexible deployment and management of honeysites.
- Honeysites are server instances, populated and distributed around the world by scripts, running real web applications and equipped with 3 levels of fingerprinting techniques:
  - behavioral fingerprinting
  - browser fingerprinting
  - TLS fingerprinting
- We registered 100 domains and ensured they were never registered before. Each domain was never advertised to users and resolved to a honeysite. Therefore, by definition, any traffic that these domains receive must belong to a bot.



Honeysites opportunistically fingerprint connecting clients across multiple layers of the stack

(Top) Daily number of new IP addresses
(Bottom) Daily number of received requests.



How do bot activities affect web server security?

- In a **7-months long** experiment, we captured **26.4M requests** from more than **287K IP addresses**.
- **57% bots are clearly malicious**, 1.3% bots are benign, 41.7% bots do not present either benign or malicious activity.
- While the majority of IP addresses in dataset are located in residential space, **only 13% of 76K malicious IP addresses appeared in online blocklists.**
- TLS fingerprinting shows that **97% bots are pretending to be browsers** while they are actually not.
- We observed requests that tried to **exploit five remote command execution vulnerabilities shortly after the vulnerability went public**, ranging from a few days to few hours.