AKAMAI BRIEF

Unpacking the Complex Dynamic Behind AI Models' Data Needs

The world is data hungry

The world is hungry for data, whether to drive business decisions, train large language models (LLMs) to develop new applications, make information on various topics more accessible and available, or allow academics to study how society evolves, human behavior, or topics as complex as climate change. Artificial intelligence models process these massive amounts of data to gain valuable insight that drives strategic decisions. Today, there is no better place than the internet to find the data. Although individuals can easily consult websites that hold the data for their analysis, collecting data manually is time-consuming and tedious. That's where scraping botnets to facilitate the data collection, and AI models to analyze the data, comes in.

Botnets are at the heart of the data collection strategy to feed AI systems. They can be programmed to scrape websites on the internet and collect key data points. For competitive analysis, the data of interest is product details, prices, and inventory from ecommerce websites. For training an LLM model designed to provide answers on many topics, the model needs verified information found in reliable media known to produce high-quality reports and articles. To assist in software development to solve various problems, the model must be trained with code examples coming from various open-source repositories.

With the emergence of AI agents such as Copilot, ChatGPT, and Claude, data collection and usage are at the top of web security practitioners' minds. Considering the relatively recent awareness of data collection needs from the broader public, there is currently a lack of understanding of how the data is collected, how it's used, and the potential long-term negative or positive impact it may have on the data owner's business. To make matters worse, this rapidly rising AI technology is only well understood by a small technology community. The space still lacks strong legal oversight to prevent misuse of the information collected. This results in most website owners rejecting the idea of sharing their data with companies running big AI models.

In this report, we focus on the various tools available to website owners to manage the data collection, the overall effectiveness of the models, and the evolving dynamics between the content owners and entities that need the data.

Website owners' unease

Beyond the uncertainty of how the collected data is used, the activity represents an operational challenge for website owners.

- **Increased operation cost:** Scraping has a price. It represents, on average, 42% of the overall site activity. In the most extreme case, scraping represents more than 90% of overall traffic volume, so legitimate user traffic represents just a small fraction (source: The Reign of Botnets by David Sénécal, 2024). Processing the extra traffic requires scaling the infrastructure up or down as the scraping activity comes and goes. It also increases the cost of delivering the content (CDN cost).
- **Site stability:** Poorly calibrated and overly aggressive scraping activity can result in site stability and availability issues, leading to revenue loss. All too often, scrapers want to get the data fast and may crawl hundreds of thousands of product pages in a short time. Sometimes, multiple scraping services attempt to collect data simultaneously, and the combined activity can easily overwhelm website infrastructure.
- **Metrics skew:** Scrapers make themselves difficult to detect. The excess traffic incorrectly identified as human by bot management solutions skews KPI metrics such as conversation rate, which most marketing teams use to decide their product positioning, marketing strategy, and advertisement investment.

Most websites' acceptable use policy specifically forbids scraping activity. However, most forums or websites supporting scraping argue that the activity is legal, and indeed, there are few laws prohibiting it. Pro-scrapers always mention the benefits that come from data collection, such as how it can help academic research, or how data used for competitive analysis can result in better pricing for the consumer.

Botnets are at the center of the data collection strategy, feeding AI models

Some people may think that very complex botnets can use AI to adapt to any bot management system dynamically. Although botnets have become more complicated as a direct effect of the advancement in bot management technology, the reality is a lot more mundane. Several bot management products that emerged in the market within the past 10 years have become a staple in the cybersecurity defense strategy of most companies. Bot management products generally identify common web search engines, bots that assist with ad serving and SEO strategy, or bots related to social media sites. Website owners also commonly use the robots.txt method to communicate with bots about which part of the site is off-limits and which bot is allowed to access the content — something that traditional "known bots" would honor.

Traditional bots provide clear intent, making it fairly easy for website owners to decide whether to allow the traffic. However, the intent and benefits of bots supporting data collection for AI agents are not always clearly understood. Until now, they have often been considered a nuisance by some due

to the excessive amount of combined traffic they sometimes generate and the fact that they don't always obey the robots.txt directive. AI platforms sometimes define their own interpretation of the robots.txt directive, as is the case for Perplexity. Here's an abstract from the page that describes how they follow the robots.txt directives: "Perplexity respects robots.txt directives. Our crawler, PerplexityBot, will not index the full or partial text content of any site that disallows it via robots.txt. However, if a page is blocked, we may still index the domain, headline, and a brief factual summary." So they won't use the data for their model, but they'll still scrape the site. As far as the website owners are concerned, they still have to deal with the traffic.

For the most part, AI agents identify themselves in the user-agent HTTP header, for example:

Company	User Agent
OpenAI	`Mozilla/5.0 AppleWebKit/537.36 (KHTML, like Gecko; compatible; GPTBot/1.2; +https://openai.com/gptbot)`
Perplexity	`Mozilla/5.0 AppleWebKit/537.36 (KHTML, like Gecko; compatible; PerplexityBot/1.0; +https://perplexity.ai/perplexitybot)`
Anthropic	`Mozilla/5.0 AppleWebKit/537.36 (KHTML, like Gecko; compatible; ClaudeBot/1.0; +claudebot@anthropic.com)`

These companies specialize in processing data through LLM models and operate simple botnets to scrape sites of interest. However, when access to the data is blocked based on their recognizable user agents, and they need the data for the success of their operation, they may resort to two strategies:

- They may contact the company that owns the data and draft a license agreement that authorizes the collection and use of the data. These license deals are often brokered with media companies.
- If no license agreement can be reached, they may outsource the data collection to companies specializing in web data collection or scraping services.

The retail and travel industry is also a big consumer of web data, and it uses AI models to extract valuable insights. This helps companies keep track of their competition and adjust their product strategy to better serve their customer base. Retail and travel companies rarely run their own data collection. Instead, they outsource the task to web data collection companies.

The web data collection industry

The web data collection industry, sometimes referred to as *scraping as a service*, specializes in collecting data on the internet at scale. Some entities have organized around the Ethical Web Data Collection Initiative or the cross-industry Alliance for Responsible Data Collection. Both organizations promote responsible and ethical data collection practices. Among other things, ethical scraping means staying away from private data, scraping responsibly by not impacting the

performance and availability of the targeted web server, or scraping during “off hours” to avoid affecting the user experience. According to G2, a peer-to-peer review site focusing on business software, “Businesses can leverage data extraction services providers to help generate leads, gather relevant information from competing business’ web pages, identify trends from document collections, and improve analysis of otherwise unstructured information.”

There are dozens of companies in that space, providing various levels of services:

- Some mainly offer infrastructure as an extended network of proxies that may include data centers, and residential and mobile IP addresses. Proxy services can be easily plugged into any homegrown scraping solution.
- Others offer scraping services with automated data extraction on top of their proxy infrastructure. They clean and structure the data to make it easier to consume, then deliver it to the customer’s data science team.
- Finally, the most advanced offering also includes extracting business intelligence from the data collected to help drive business decisions.

Customers of these services can define their targets, the frequency of the data collection, and the level of service they want. Web security practitioners and web security solutions vendors face a technology-advanced adversary staffed with a team of engineers and data scientists, which evens the odds against the team of engineers and data scientists that build bot management products. As bot detection technology advances, scraping technology must also advance to continue collecting data and protect revenue.

Data collection resilient to bot detection

Like bot management solutions, which have complex workflows to detect bots, advanced scraper solutions have advanced workflows to detect bot management solutions to ensure the success of their data collection. At a high level, the workflow may look like the one in Figure 1. The botnet first detects any protocol-level fingerprinting (TLS, HTTP) detection that exists and adjusts its fingerprint as needed. When the data collection is still unsuccessful, the botnet also looks to change its load balancing strategy to defeat IP reputation or IP-based rate-limiting detection methods. Finally, the botnet checks for any JavaScript (JS) fingerprinting:

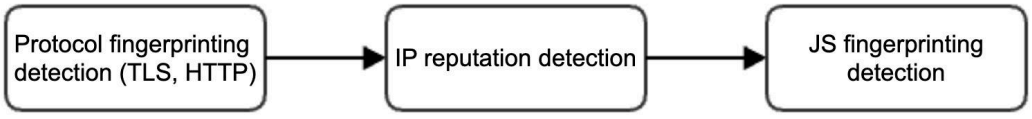

Fig. 1: High-level anti-ban detection logic

Protocol fingerprinting detection

Most bot management solutions evaluate the HTTP headers. Some even assess the TLS handshake when the secure connection is established. As seen in Figure 2, to defeat the bot detection, the scraper will need to:

1. Evaluate the set of cookies and assess which one to ingest and replay to collect data from the site successfully.

2. Adjust the HTTP header set sent, its order, and its values to match the system it pretends to be (e.g., Chrome on Windows, Firefox on MacOS, Edge on Windows, Safari on iPhone).
3. Adjust their TLS handshake settings to match the OS and browser they claim to be.

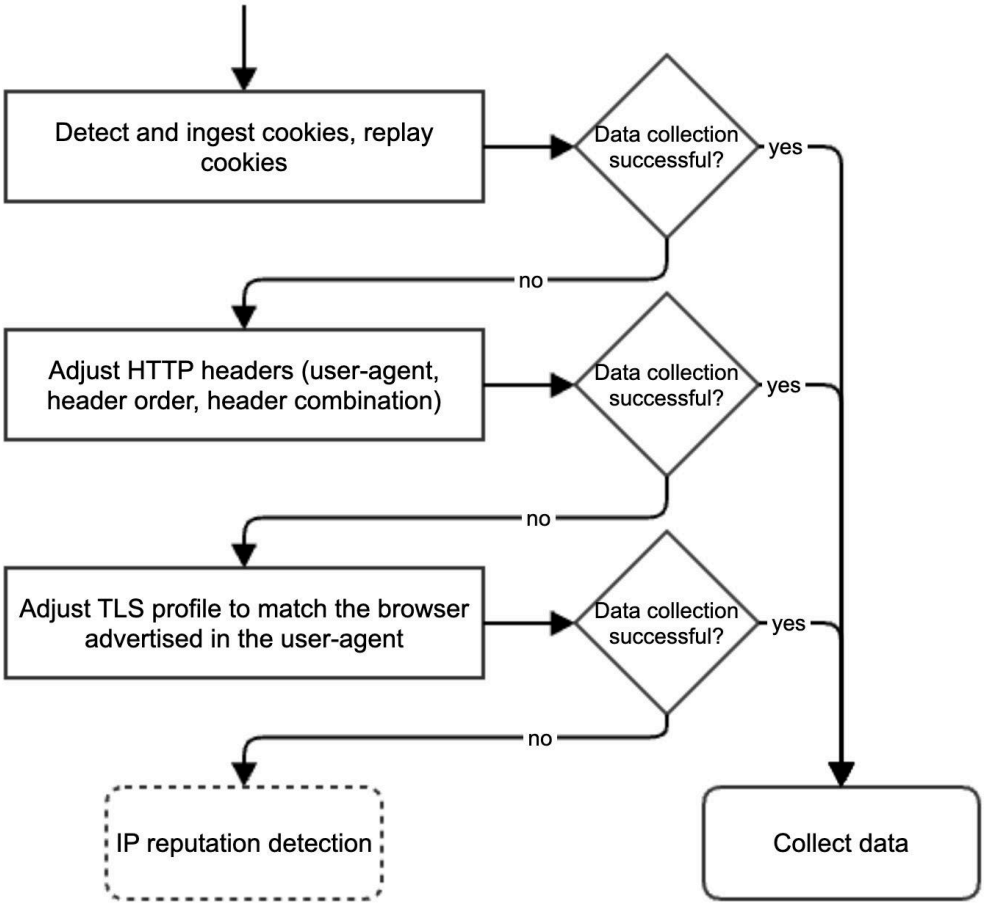

Fig. 2: Protocol fingerprinting detection

IP reputation and rate limiting detection

To defeat IP reputation systems, the botnet must load balance the traffic through a large set of IP addresses. The cost of the proxy service depends on the type of IP used. A proxy service consisting of IP addresses hosted by cloud providers is significantly cheaper than the ones that include residential and mobile IP addresses. When choosing the correct type of proxy to use, the botnet will follow this logic:

1. Start with cheaper proxy servers hosted by cloud providers.
2. Cloud providers are easy to detect and may be blocked. In case of excessive data collection failure, the botnet may be adjusted to use more expensive proxy services that offer residential IP addresses.
3. Further adjustments may be required, introducing even more expensive mobile IP addresses and removing the cloud provider to defeat detection.

The number of IP addresses used varies depending on the quantity of data to collect. It is very common to see scraping activity from a single botnet using more than 500,000 unique IP

addresses. The transition from cloud provider IP addresses to the more advanced and expensive solution is based on the data collection success rate. Figure 3 shows the logical progression.

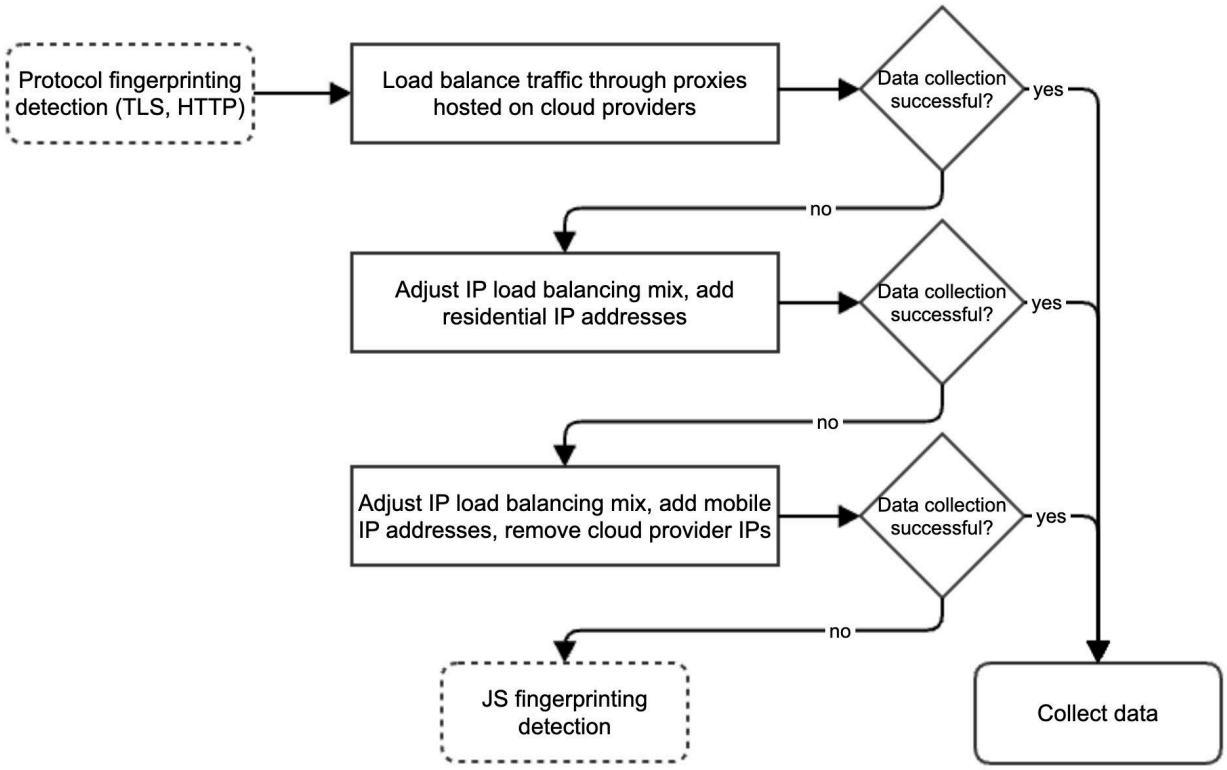

Fig. 3: IP reputation and rate limiting detection

The botnet also looks for known JavaScript and data collection associated with common bot management providers. Depending on the complexity of the solution and JavaScript, the botnet may need to execute it to counteract these methods. Again, the botnet will attempt to run the cheapest solution that yields a good success rate. As seen in Figure 4:

1. The scraper may try to harvest good fingerprints from legitimate user interaction and replay them.
2. If this simple trick is insufficient, running the bot management vendor's JavaScript through a minimal JS execution engine like Js2Py or equivalent frameworks may be necessary. However, this may not be enough if the site is protected with a more advanced bot management solution. At this stage, the botnet operator may attempt to reverse engineer the JavaScript to develop a more advanced script to simulate the JS execution.
3. Running the botnet on a headless browser may be necessary if the previous steps don't work.
4. Tuning the headless settings for stealth mode may be required to avoid detection by advanced bot management solutions.

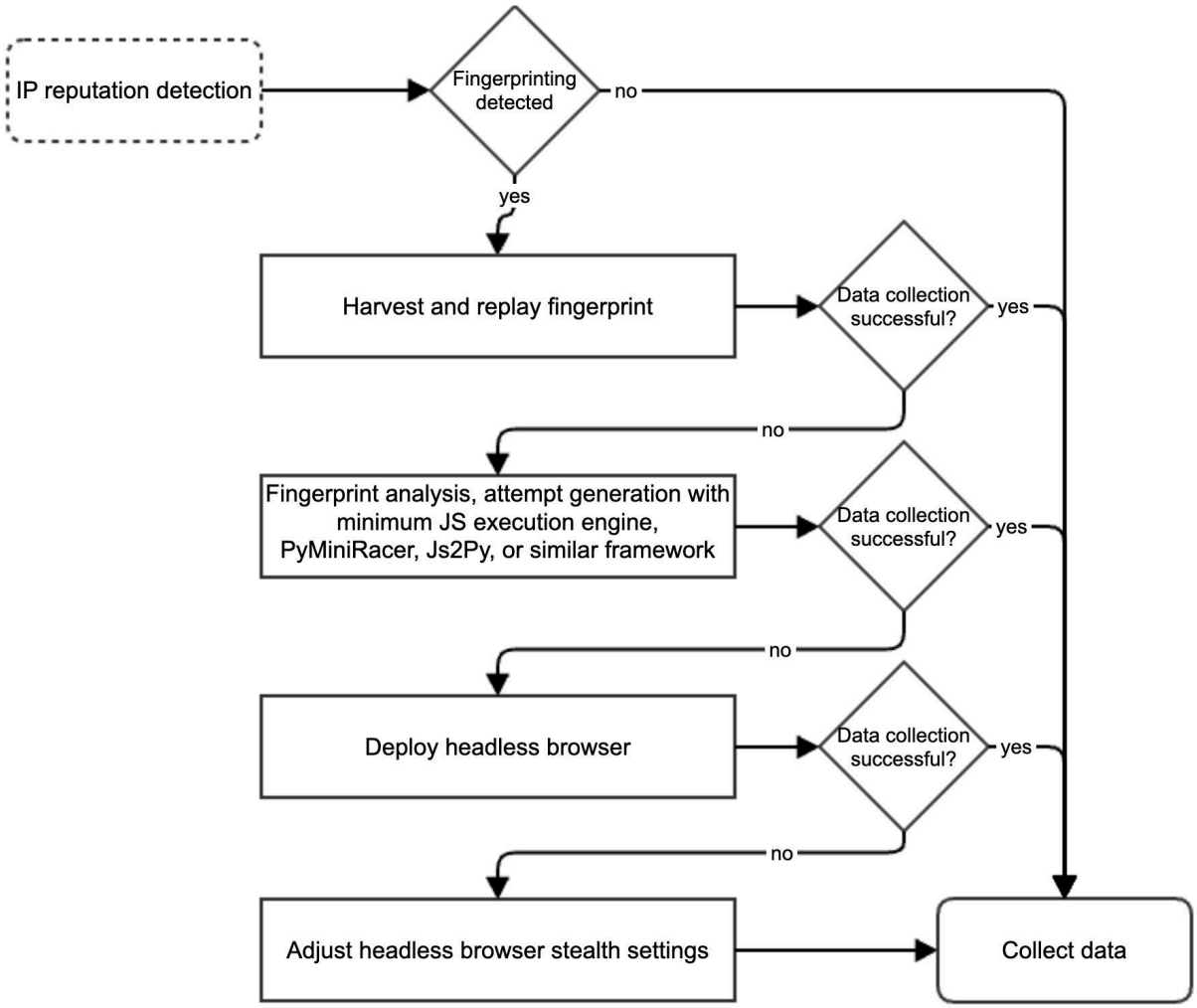

Fig. 4: JS fingerprinting detection

Some of the above workflows may involve machine learning, especially when evaluating the JavaScript fingerprinting code, adjusting to the cookie strategy, HTTP header, and TLS parameters, and collecting statistics on the success rate. This logic is continuously running, and the anti-bot strategy and configuration must be adjusted when the data collection success rate drops.

How scraping platforms acquire residential and mobile IP addresses

As previously discussed, a network of proxies is a common and required feature of scraping services. This is key to their traffic load balancing strategy to defeat detection centered around IP reputation, IP blocking, and IP rate limiting. Basic proxy services are deployed on virtual machines in data centers and are, as such, easier to detect. However, the most advanced proxy services offer residential and mobile IP addresses. All the web data collection services that belong to the alliances

mentioned earlier claim ethical sourcing of IP addresses. Here are a couple of models to source IP addresses:

- **Free mobile app monetization incentive:** Scraping platforms offer a mobile app software development kit (SDK) as a means to monetize free applications (for example, games), increase revenue, enhance the user experience, or provide an ad-free experience. Bright Data (Bright SDK), Packetshare (Packet SDK), PYPROXY, Infatica, and Proxyrack are a few examples of companies that offer an SDK that, when integrated into an application and enabled by the user, will make the device become part of a proxy network. Once a device opts into the proxy service, some web scraping activity may transit through the device based on demand. The SDK provider pays the mobile app developer based on the active daily users that opt in to the proxy service.
- **Bandwidth monetization:** The companies Honeygain, PacketStream, Pawns.app, and Packetshare offer an application that allows users to earn passive income by sharing their internet bandwidth and participating in the proxy service. Users get paid based on the amount of traffic that transits through their devices. The application runs in the background and doesn't require any user intervention.
- **Free VPN offering:** Bright VPN offers a free VPN service in exchange for users sharing their bandwidth and opting into Bright Data's proxy network.

The user must consent to become part of the proxy service and share their bandwidth before taking advantage of the offering. This is, as such, an ethical sourcing of IP addresses. These services are advertised to the user as a way to help the internet and the data science community. Although these services help researchers in their important work, the primary source of revenue for these web data collection companies is the collection of product pricing and inventory, such as airfare or hotel bed availability. Companies need near–real time data updates to make strategic decisions on product or service positioning and pricing to attract consumers and increase revenue.

An emerging pay-as-you-scrape model

Everyone wants data from others, but no one wants to share their data with others. What if there was a better way? Recently, we've seen the emergence of a handful of start-ups like TollBit and Skyfire that offer a payment gateway for content owners to monetize the scraping activity. If one can't stop it, one might as well get paid for it. As seen in Figure 5, these payment gateways partner with bot management vendors to do the detection. This seems like an interesting approach. However, a few questions remain unanswered:

- **Fair price:** What represents a fair price for the data from both sides? How much is the party collecting the data willing to pay, and how much is the content owner expecting to receive for the content? Are some content categories more valuable than others, and can a fair price be agreed upon between both parties involved?
- **Respectful data collection:** Will the data collection vendors play fair? Is there some entirely off-limits content, and will the web data collection vendor honor this, or still attempt to get the data no matter what if their customer needs it?
- **Transparency on data utilization:** Can the data owner still have some oversight into how their data is used once collected to ensure it will not negatively affect their business?

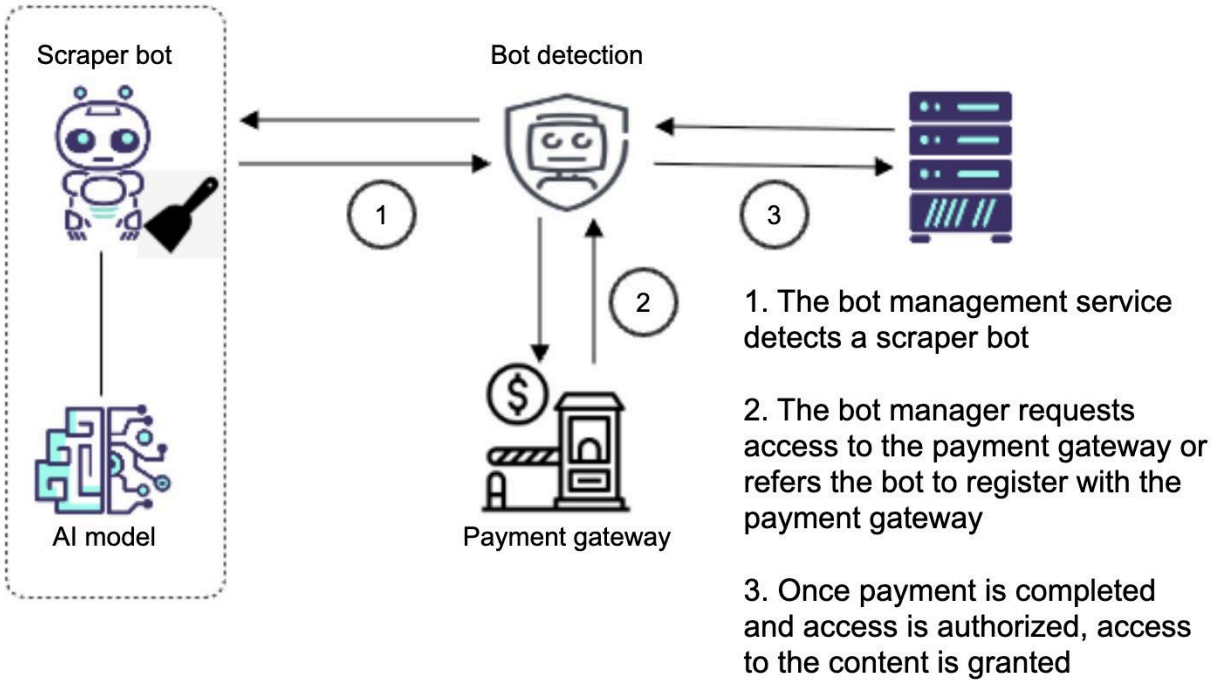

Fig. 5: Paid access to data

For media companies, getting paid for the scraped content is not as important as getting attribution for the content and enabling the connection with the content publisher and its author for recognition, subscription, and ad revenue. The start-up ProRata.ai offers a service to help with the attribution of the content collected.

If this model takes off, it could considerably change how data is collected online, which is needed to bring a level of sanity into the excessive scraping activity. However, for this model to work, it will require a lot of education, a general mindset shift on the scraping problem, and strong collaboration between the various parties.

The role of bot management products

Bot management products play an essential role in this ecosystem, giving website owners a set of detection methods to effectively detect, categorize, and observe botnets collecting their data. They also provide a set of response strategies to manage the activity. Bot management products must adapt to the complex logic described previously and not rely on the bot operator's goodwill to clearly identify themselves in the request (typically, the HTTP user-agent header) or to diligently follow the directives from the robots.txt. The traditional choice of response so far has been to monitor, deny, tarpit, delay, or even challenge the detected bot traffic. In the future, if the pay-as-you-scrape model gains traction, toll-based may become an alternative to existing response strategies.

As we've seen, AI agent scrapers and scraper services rarely follow the robots.txt directives, frustrating content owners. In response, the concept of an "AI labyrinth" emerged, which is like a more modern version of the honeypot. It consists of using LLMs to generate fake data that looks similar to the original data but has inaccurate information. Some variants dynamically generate fake

links to keep sending the botnet on a never-ending link chase that doesn't lead anywhere. Bots detected by the bot management solution are served alternate content to send them to the AI labyrinth. The fake content is intended to pollute AI models that use it. Although innovative, the fake content must be realistic to be convincing and effective in the long run. This requires regular training of the model, maintaining a minimal infrastructure to generate and serve the content, which can become expensive to maintain and operate over time. Also, bot operators may get fooled initially, but will quickly learn to recognize the signs of the AI labyrinth and adjust the characteristics of their botnets as previously described to defeat detection and regain access to the original site content.

The economy that AI models' data addiction supports

The technology involved is complex, but the interconnection between the various industries and their market size is nontrivial. Grand View Research estimates that the AI agent market was worth $5.4 billion in 2024. Although the media industry focused on their data needs in recent months, the retail and travel industry is likely the biggest data consumer today. Some may attempt to build their in-house scraping infrastructure, but may outsource the data extraction to data collection companies when they encounter too much resistance from bot management products protecting the site.

According to Forbes, the revenue of the web scraping software market was estimated at $800 million in 2024. Web scraping companies have built and operate vast proxy networks whose success relies on acquiring millions of residential and mobile IP addresses through their SDK offering, which provides game developers with an opportunity to increase their revenue and gives gamers a chance to get an enhanced experience. Estimating that the web data collection industry spends 5% of its revenue to acquire residential and mobile IP addresses, it contributes approximately $40 million to the mobile app gaming market.

Companies targeted by the scraping activity protect their assets using bot management products, a growing industry about the same size as the web scraping software industry, which reached about $732 million in revenue in 2024, according to Grand View Research. The nascent web scraping monetization service does not show significant traction at this time, and no meaningful data is available to estimate its revenue.

In the end, the retail, travel, and AI agent industries generate revenue for the web data collection industry and indirectly generate revenue for game developers. The retail and travel industries also generate revenue for the bot management industry. The web scraping monetization industry may eventually get its revenue from the web data collection platform and AI agent, and share a fraction of its income with the bot management industry and the retail and travel content owners. Figure 6 illustrates the connection between the different industries.

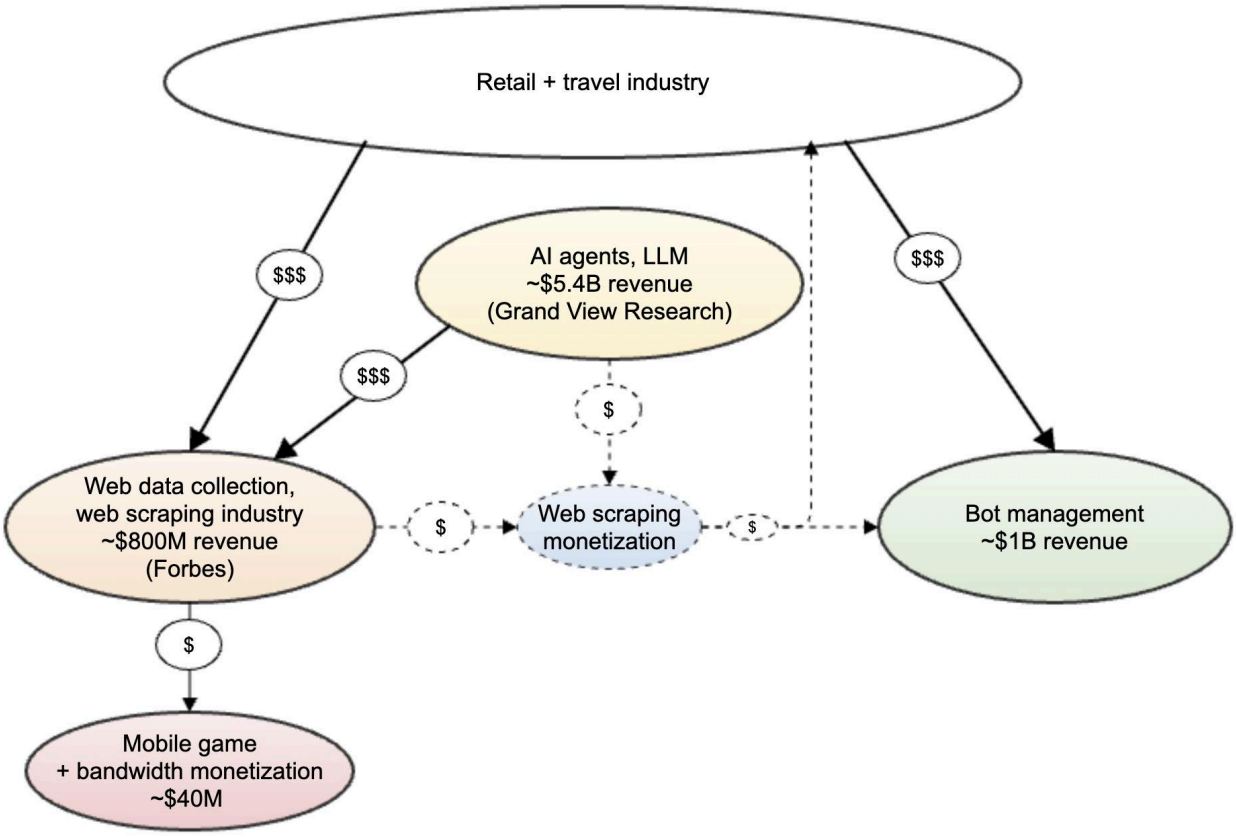

Fig. 6: The connection between companies running AI models and adjacent industries

Conclusion

The internet ecosystem and the applications built around it continuously evolve, and the web data collection and bot management markets that serve opposite needs are colliding. Both have their relevance and purpose. This tug of war remains unresolved partly because website owners are ambivalent about their need for data and willingness to share their own. The emerging monetization service could bring some sanity into the ecosystem. Imagine a world where one could clearly identify the bot activity and its intent, make an informed decision to share and monetize their data, and more easily clean up their metrics. This more open data market would require collaboration, coordination, and transparency between all parties involved. It could actually help reduce bot traffic on the internet. If web data collection companies don't struggle as much to collect the data, the number of attempts required — and consequently the number of requests — will go down.