BOTS Over the Internet

Priti Rani Rajvanshi[1]
[1]Assistant Professor (School of IT)
Institute of Management Studies Noida,
Noida, U.P., India

Osho Bhadani[2]
Student, MCA IMS, Noida
Noida, U.P., India

Pragati Bisht[3]
Student, MCA IMS, Noida
Noida, U.P., India

***Abstract*:- Our life majorly depends on a virtual world, where the intensity of interacting with machines is higher than human networking. In this internet age, bots over the internet have created a sensation. These bots are used virtually everywhere by every major companies in the world and are taking Artificial Intelligence to an altogether new level.**
As this technology is still taking shape, most bots follow a set of rules programmed by a human from a bot building and creating platform. While bots may be used for productive and informative tasks, despite their neutral origins, they often come in the form of malware.

Determining crucial web traffic trends and making decisions based on distorted data can be really challenging. As bots and ad frauds are the real enemy of not only the advertising industries but also any company with a significant online presence. So, what exactly are bots?

Keywords:- Bot, Internet bot, web crawler

1. INTRODUCTION

Bots are basically a set of algorithms that are designed to perform automated tasks which require running scripts over the internet. These applications perform simple and repetitive tasks efficiently and quickly, with greater expertise than any human. Bots are mostly utilised to do laborious and monotonous tasks, but can also be used in malicious criminal activities which can be done online. These bots can be called as virtual robot, an equivalent part (clone) of their physical versions without a mechanical body. Bots are largely used in web spidering (web crawler), in which an automated script fetches, analyses and files information from web servers at many times

the speed of a human can do. More than half of all the web traffic is made up of bots.

Fig. 1: Bots connections.

Search engines uses bots to surf the web and orderly and systematic catalogue information from websites, trading sites make them look for the best bargains in seconds, and some websites and services employ them to deliver important information and notification like weather conditions, news and sports, currency exchange rates.

Bots are foremost used by search engines like Google, Bing, Yandex (Russian search engine) or Baidu (Chinese search engine) for web spidering purposes. These bots collect information automatically on regular basis from hundreds of millions of domains and index it into their result pages.

Fig. 2: Crawler.

Unfortunately, all bots roaming the internet are not useful and harmless. Blackhats (cyber crooks) have also noticed their potential and have come up with malicious bots – programmed and designed to secretly install themselves on unprotected or vulnerable computers without the knowledge of the user and carry out whatever actions they demand. And that could be anything from data theft, sniffing or sending spam to participating in a distributed denial of service attack (DDoS) that brings down entire websites.
Around half of web traffic are bots.

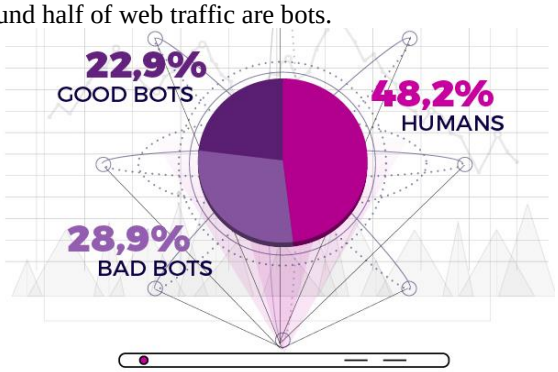

Fig. 3: Web traffic breakdown.

2.GROWTH OF BOTS

Bad bots among all website traffic in 2018	20.4%
Growth in bad bot traffic from previous year	6.4%
Good bot traffic percentage in 2018	17.5%
Growth in good bot traffic from previous year	-14.4%
Human website traffic percentage in 2018	62.1%
Decrease in human traffic from previous year	+7.5%

Fig. 4: Traffic of bots vs Humans.

The current generation of bad bots are described as Advanced Persistent Bots (APBs), which have characteristics that make it difficult to alleviate against. Advanced Persistent Bots (APBs) try to obfuscate their origin by relying upon different unknown proxies and other identity-hiding technologies, while simultaneously trying to appear to target sites as legitimate human traffic.

As per Distil survey, 49.9 percent of all internet bots appear as browsers running Google Chrome. And another 28.2 percent masquerade as other popular browsers, along with Firefox, Internet Explorer, and Safari.

Astonishingly, 53.4 percent of all bad bot traffic comes from the USA, with the Netherlands, with 5.7 percent of all traffic, being the second most common country-of-origin. This is likely due to the large number of data centres and hosting providers in both countries.

On the other hand, the most commonly IP-blocked countries are Russia and the Ukraine, highlighting the huge disparity in where people perceive attacks to originate from, rather than where they actually do.

1	United States	53.4%
2	Netherlands	5.7%
3	China	3.9%
4	Germany	3.9%
5	Canada	3.2%

Fig. 5: Top 5 Bad bot traffic by country.

1	Russia	32.6%
2	Ukraine	15.6%
3	India	15.2%
4	China	11.2%
5	United States	6.6%

Fig. 6: Top 5 most blocked country.

3. GOOD BOTS VS BAD BOTS

In general, bots can be broken down into two categories – the good bots and the bad ones. Good bots are created to make humans lives easier and their activities involve: web crawling, website monitoring, the content retrieving, data aggregation, online transactions and so on. The bad bots bring fake (sham) traffic to the website they are programmed for and their malicious intent may involve: stealing valuable data, content/price scraping, posting spam comments and phishing links, distorting web analytics and damaging SEO, contributing to distributed denial-of-service (DDoS) attacks etc.

More than 50% of bot traffic is used for the malicious purposes that we encounter on the Internet websites. Below I presented the most common types of good and bad bots, along with some examples. There is no predefined, general classification of bots and you may come across different categories – some more general, others more specific.

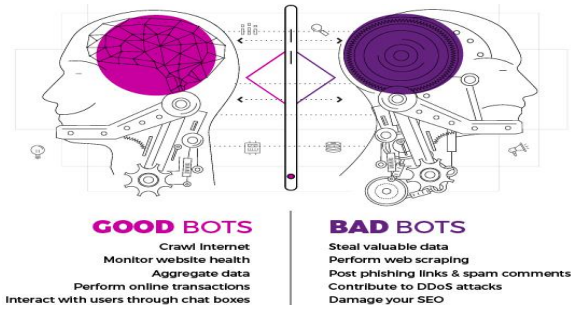

Fig. 7: What good bots & bad bots do?

The good bots:

Crawlers / spiders (e.g. Googlebot, Yandex bot, Bingbot): Used by search engines and online services to discover and index website content, making it easier for internet users to find it.

Traders (Bitcoin trading bots): Used by Ecommerce businesses to act like agents on behalf of humans, interacting with external systems to accomplish a specific transaction, moving data from one platform to another. Based on the given pricing criteria, they search for the best deals and then automatically buy or sell.

Monitoring Bots (e.g. Pingdom, Keynote): Monitor health system of the website, evaluate its accessibility and report on page load times & downtime duration, keeping it healthy and responsive.

Feedfetcher / Informational bots (e.g. Pinterest bot, Twitter bot): Collect information from different websites to keep the users or subscribers up-to-date on the news, events or blog articles. They cover different forms of content fetching, from updating weather conditions to censoring language in comments and chat rooms.

Chat bots (e.g. Messenger, Slack, Xiaoice):A service that enables interacting with a user via a chat interface regarding a number of things, ranging from functional to fun.

The bad bots:

Impersonators: Designed to mimic human behaviour to bypass the security and by following offsite commands, steal or bring down the website. This category also includes propaganda bots, used to manipulate public opinion.

Scrapers: Scrape and steal original content and useful & relevant information. Generally repost it on different websites. Scrapers can reverse-engineer pricing, product

catalogues and business models or steal customers lists and details for spam purposes.

Spammers: They post phishing websites links and low-quality promotional content to lure visitors away from the website and ultimately drive traffic to the spammer's website. Often use malware or black hat Search Engine Optimization (SEO) techniques that may lead to blacklisting the infected site. A specific type of spammer is auto-refresh bots, which generate fake and bogus traffic.

Click & download bots: Intentionally interact or click on Pay per Click (PPC) and performance-based ads. Associated costs of such ads increase based on exposure to an ad – meaning the more people are reached, the more expensive they are. This form of ad fraud is new, but already common among bots developers. According to paid advertising experts, one in five paid clicks were fraudulent.

To give you an idea of all the good bots and bad bots out there, here's a list of some of the most common bots you'll find accessing your site:

Good Bots:

- Monitoring bots – 1.2%
  - Website health checkers
- Commercial crawlers – 2.9%
  - Metric crawlers (AHREFs, Majestic)
- Search engine bots – 6.6%
  - Googlebot and Yahoobot
- Feed fetchers – 12.2%
  - Bots that convert sites to mobile content

Bad Bots:

- Impersonators – 24.3%
  - Bots that look real and are often used in DDOS attacks
- Web scrapers – 1.7%
  - Used to scrape prices and content off websites
- Spambots – 0.3%
  - Used to post silly comments and messages
- Hacker tools – 2.6%
  - Bots that scan for vulnerabilities on websites

How bot traffic damages performance?

There are number of ways how bots can affect your webpage and your business's performance.

Bots contribute to DDoS attacks: A DDoS attack (distributed denial-of-service) is a malicious attempt, which makes a server or a network resource unavailable to users. DDoS attacks are generally performed by botnets – a group of hijacked Internet-connected devices, injected with malware. Botnets are controlled from a remote location without the knowledge of the device's owner. A successful DDoS attack results not only in short-term loss but can have long-term effects on your online brand reputation, generate significant costs from hosting providers or even compromise your business.

Damage your SEO and website reputation: Firstly, scrapers stealing your content and illegally distributing it on other websites might degrade your Search Engine Optimization (SEO) and outrank you on search engine listings. Secondly, if your website will get a lot of fake visits, views or comments generated by malicious bots, search engines will undermine its credibility. Since advertising networks consider fake views as a form of fraud, you might end up with a penalty on your website. If this happen again & again, advertising networks could even blacklist or remove your website.

Bots can take over your account: Bots can hack your website, steal your data and make it available on hacker dump sites and black markets. Loss of customer's sensitive details & information can impact brand reputation greatly and result in high costs.

Bots can lead to monetary loss: Besides all the threats listed above, bots can lead to direct monetary loss – Your paid ad campaigns will be more expensive and less effective because of the fraudulent clicks; your visitors might be lured away from your site via comment spam links and poor UX; stolen content might require high-priced legal actions. Not to mention that the server and bandwidth cost increase rapidly when bots hit the website with millions of unwanted requests within a short time frame.

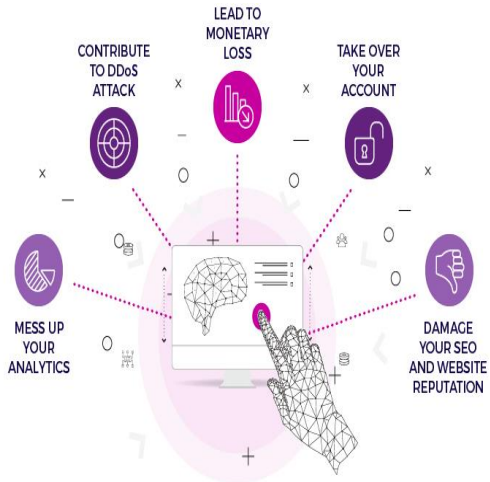

Fig. 8: How can bots damage any websites?

How to determine bots activity?

If you want to determine if bots are messing with your webpage, you need to dig a little into your analytics. What can indicate bot's suspicious activity?

Uneven traffic: Unusual increase in your page views and you haven't recently run a big ad campaign, bots may be standing behind it.

Abnormally low time spent on a page & increased bounce rates: As bots are programmed to perform their tasks at high speeds, they can crawl number of pages within

a small time frame. If you see many page durations which only span a few seconds, then you might be looking at a bot's activity.

More visits than actual customers: If you're noticing a sudden increase in your monthly website visits, you should check where your traffic is coming from.

Unknown domains referring traffic to your site: If you suddenly started to see a spike in referral traffic or a large number of users checking your site directly every day, it's probably robots.

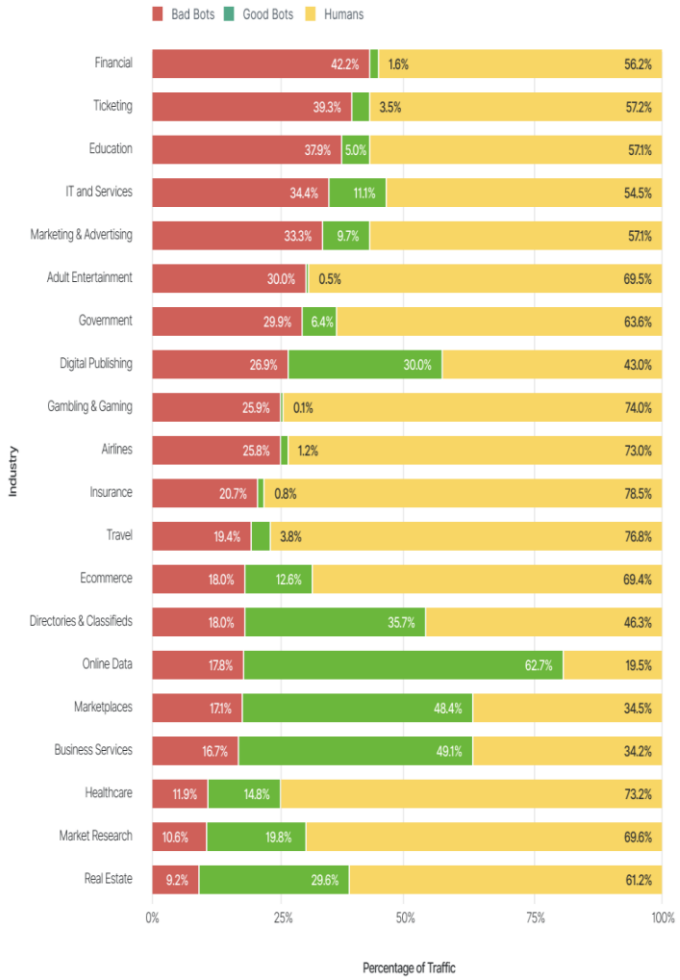

Fig. 9: Bad Bots Vs Good Bots Vs Human Traffic – By Industry.

4. BOTS LIFECYCLE

Bots are exciting. Everyone is talking about it and everyone wants to build it. Some of the early bots have gotten slightly mixed reviews — they're interesting, but they're essentially a play thing on a system. As people start working on bots with more capabilities, developers must appreciate and plan for the entire bot lifecycle.

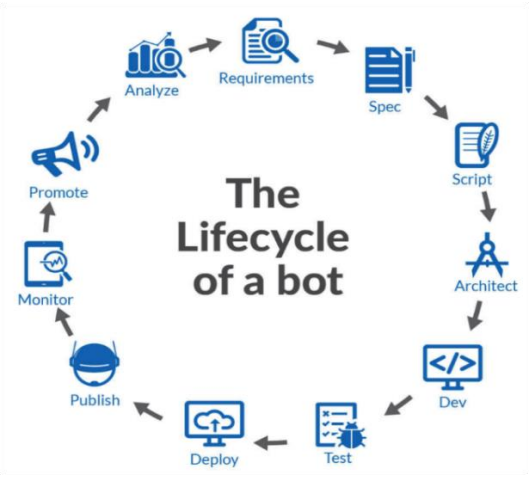

Fig. 10:Steps of a bot life cycle.

a. Requirements

Gather market requirements for the bot: who's the target customer, what are the main points and what benefits the solution will deliver. This initial step is similar to any other software project, though other steps below are unique to bots.

b. Spec

Develop a product specification for the bot identifying the features and functionality of the bot. These features should deliver the benefits identified in the Requirements step. Note that the specification must also include short and long description of the bot along with other collateral that will be required later in the Publishing step.

c. Script

While the first 2 steps are similar to any other software lifecycles, this step is unique to the bot building process. While websites and applications have structured interfaces, bots have a conversational interface. Instead of building wireframes like we do for websites and applications, this involves building conversational scripts that represent user interactions.

The conversational script must be representative of actual human conversations. Since the conversational interface has no tabs or prompts, the bot script must guide the user towards fulfilling the desired task.

d. Architect

Create the engineering design for the bot. This includes both the front-end and back-end components. The front-end is the conversational interface — translating user input into specific actions and vice versa. The back-end is the computation performed by the bot as well as integrations to other web services.

e. Dev

This stage is the development stage where the bot is developed. Given the conversational interface, bot developers generally repeatedly jumps a lot more between coding and testing than in traditional software development. When the bot is coded to handle a specific set of

conversational statements, it's a good practice to unit-test the code through the messaging interface.
Developers must also insert tracking probes (a program for monitoring & collecting data) into the bot — these will be helpful at the Track stage.

f. Test
As described above, the testing is deeply intertwined with the development process. However, testing is tricky for a bot developer. The code must be tested in emulator, as well as in the actual messaging platform too. Since the different types of messaging apps, and the differences in message rendering, this can be a time-consuming process. Also, different messaging platforms have different rules and guidelines and access limitations for test bots.
Apart from the unit-testing during the development cycle, this step also includes the final Quality Analysis of the bot. The Quality Analysis process must run the bot through the conversational scripts developed above. The Quality Analysis process also needs to be aware of, and ensure compliance with, the Publish guidelines of the messaging platform. Messaging platforms have different guidelines. (e.g. bots must not spam, they must introduce themselves, explain themselves, behave themselves, handle exceptions etc.).

g. Deploy
Once the bot is built, it must be deployed to a hosted platform. The hosted platform must be stable and needs its own monitoring and devops support.

h. Publish
Once the bot is tested and deployed, it must now be submitted to the various application stores for approval. Each messaging platform has a different approval process, with varying different methods. When submit it require a variety of descriptive elements such as a short description, a long description, images, scripts, videos etc. (best way is to develop these in the Specification phase).

i. Monitor
Once the bot is published, it must be monitored regularly. It doesn't mean just ops monitoring as mentioned in the Deploy phase. It means monitoring the bot using actual conversational scripts. The ops monitoring may indicate systems are well, but the bot may still be unresponsive to any user conversations. The best way to monitor the bots by monitoring the user conversation.

j. Promote
While publishing to the bot store is the first step, there are other options to make your bot discoverable. There are cross-platform bot stores that are off the messaging platforms that may still drive traffic to your bot.

k. Analyse
As the bot starts being used, its performance must be tracked and results analysed. This involves reviewing both conversation logs and usage metrics. Analysing conversational behaviour is different from analysing click-paths for websites and apps.

Repeat
The learnings from the Analyse phase can be cycled back into the bot development process to build an ever-improving bot. Some bots may even be programmed with self-learning AI programs that need inputs from users and trainers to continuously improve themselves.
Building good bots isn't easy. A little bit of structure and process goes a long way in achieving bot success.

5. Approaches for tackling bots

At a very high-level, there are three primary ways to mitigate bots.

The static approach: The quickest way to identify and mitigate bots is by using a static analysis tool. By looking at structural web requests and the header information and co-relating that with what a bot claims to be, we can passively determine its true identity

The challenge-based approach: A more progressive way of addressing a bad bot is a challenge-based approach (or support based approach). Websites are equipped with proactive components to measure how a visitor behaves with different technologies — for instance, how it supports cookies and what cookies it supports. It also takes a close look at JavaScript, what kind of JavaScript it can run, and what objects are accessible in that JavaScript. We can also use scrambled imagery like CAPTCHA, which usually requires a more dedicated attack to bypass.

The behavioural approach: Beyond the static approach and the challenge-based approach is a behavioural approach to bot mitigation. This is where we look at the activity associated with a particular bot. In other words, a good way to uncover a bad bot is to find out what it claims to be. Most bots are likely to link themselves to a parent program like JavaScript, MSIE, or Chrome. If the bot's characteristics vary in any way from the parent program, this anomaly will help mitigate the current problem and any problems in the future.

The most effective way to identify and mitigate bots is by using a specialized tool in a combined multilateral approach.

6. RESEARCH

Visualized Bot Attacks on Healthcare Data:
Detailed threat intelligence gathered from a healthcare-specific honeypot created to uncover current attack trends, methods, and identify which data is most frequently targeted and therefore at a higher risk of being compromised.
Web application Firewalls (WAF) and other common network protection tools are important network protection but they are commonly bypassed by bots traffic that

masquerade as legitimate & valid users. Based on all previous data collected from protected websites roughly around 50% (half) of autonomous bots are able bypass Web application Firewalls (WAF). The details of this case study highlight the importance of monitoring traffic using advanced Bot tools in order to detect suspicious events and restrict all the malicious traffic.

Flow of Events

On August 28, 2019, a healthcare honeypot that included a health-related news feed and a fake customer login page (great bot bait). It took merely less than one week for bots to find and begin quietly attacking the website. In the initial three week period, it was detected and identified that more than 95% of the website traffic is due to bots:

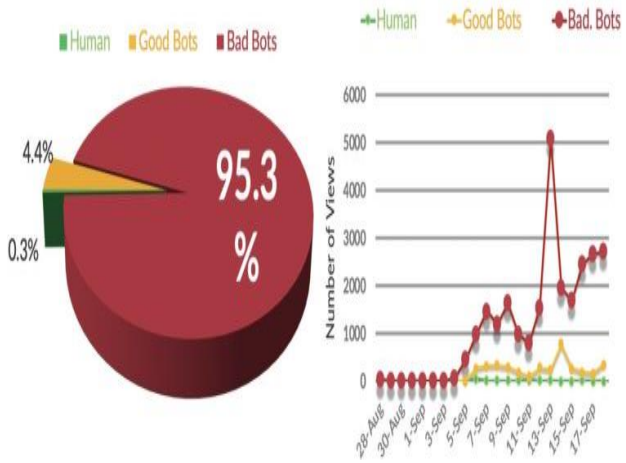

Fig. 11: Website traffic (Aug. 28 to Sept. 17).
Fig. 12: Bad Bot Trend (Aug. 28 to Sept. 17).

Bots Detected

At the end of the first week, crawlers like Ahrefs and SEMrush bots began probing (exploring or examining) the website looking for vulnerabilities and valuable content, including account login and account creation links and pages. Once the bots identified the targets, scripted bots and more complex bots like PhantomJS began to perform credential stuffing attacks on the fake customer login.

In the chart below, we can see that attackers were focused on the login.php sites. In tandem, a much smaller volume of traffic was scanning the info page and the main homepage looking for other vulnerabilities.

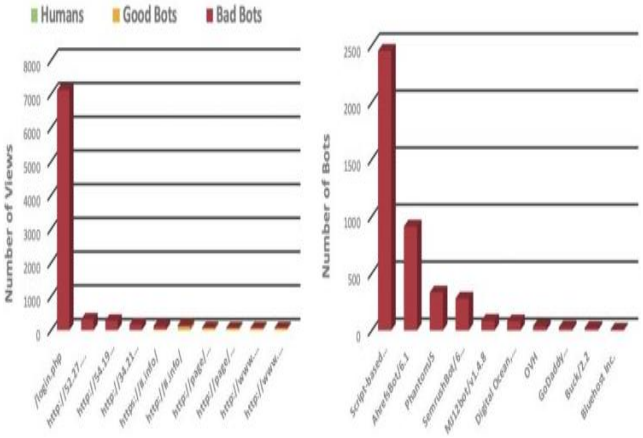

Fig. 13: Top 10 bot views by page (Aug. 28 to Sept. 17).
Fig. 14: Top bad bot types (Aug. 28 to Sept. 17).

Attack Method Details

Fraudsters performed credential stuffing (or password guessing) attacks using common login pairs like admin/password, admin1/password1. It was also recorded a large list of credentials pairs that was from a known leaked credential list.

During the attack period, the majority of credential stuffing attempts used script-based bots. When this type of attack failed, more advanced tools like PhantomJS were deployed in an attempt to identify "good" customer login credentials. For most standard security tools like WAF, PhantomJS and scripts are difficult to identify because they look like a real user. PhantomJS and similar software packages were developed as a QA tool to run full scale testing on websites to simulate large volumes of human traffic. Unfortunately, these tools have also been adopted by attackers to run large Bot attacks against websites and also deploy them using VPN to mask their location and bypass most common network detection methods.

Result

Bot-led automated attacks have increased 171% Year over Year, and 30% over the last six months, according to a recent report by LexisNexis. Simple, low cost bot attacks and automation are used by bad actors to scan huge swaths of the internet looking for vulnerable targets that are considered to be "low hanging fruit." Using a Darwinian selection process in cyberattacks, methods that are successful proliferate, while unsuccessful approaches die off quickly.

- **Global Insights from the LexisNexis Digital Identity Network (January – June 2019) Include:**

  • 277 million human-initiated cyberattacks, up 13% in just six months
  • 111 million mobile attacks, rising nearly 10% during that same timeframe

• 171% year-over-year growth in bot-based account creation attacks against retail e-commerce
• 144% surge in mobile app registration fraud across banking, media and more
• 52% jump in payment attacks in North America, compared to 12% globally

As we saw in this honeypot, the attacks started with commonly available web scanning and crawler tools to examine and probe the environment for vulnerabilities like unpatched systems. Even if no vulnerabilities were found at the start of the attack, fraudsters can still revert to account takeover techniques by testing stolen login credentials so they can gain access to additional pages of a website.
Unfortunately, web scanners and QA tools like PhantomJS are available to bad actors and CIOs/CISOs alike. Although WAF is useful for identifying these scanners and stopping known attacks like SQL Injection (SQLi), it is usually only successful against 40-50% of actual automated threats.

7. CONCLUSION

As bots cover more than half of the whole web traffic and most of them have malicious intents, it's more important than ever to protect your website from bogus traffic. Over time, the negative effects of bot traffic and ad fraud can heavily influence your web strategy, leading to wasted money, sales, time and effort.
Bad bots are responsible for a very large number of serious security threats to your website. You can harden your site security by analysing traffic for bots and identifying malicious clients, and block them preferably in a transparent manner that doesn't affect your visitors.
One way to do this is through the use of web application firewalls or application delivery controllers (Web Application Firewall (WAFs) and Application delivery controllers (ADCs)).

REFERENCES

[1] "Bot Traffic", Website:
[2] https://voluum.com/blog/bot-traffic-bigger-than-human-make-sure-they-dont-affect-you/
[3] "Bot Lifecycle", Website:
[4] https://chatbotsmagazine.com/the-bot-lifecycle-1ff357430db7
[5] "Growth of bots", Website:
[6] https://thenextweb.com/security/2019/04/17/bots-drove-nearly-40-of-internet-traffic-last-year-and-the-naughty-ones-are-getting-smarter/
[7] "Good bots vs Bad bots (By Industries)", Website:
[8] https://www.zdnet.com/article/bad-bots-focus-on-financial-targets-make-up-20-percent-of-web-traffic/
[9] "What percentage of bots are good bots & bad bots", Website:
[10] https://ppcprotect.com/how-many-of-the-internets-users-are-robots/
[11] "Approaches for tackling bots", Website:
[12] https://www.imperva.com/blog/understanding-bots-and-your-business/
[13] "Visualized Bot Attacks on Healthcare Data", Website:
[14] https://www.botrx.com/blog/botrx-intelligence-report-how-we-visualized-bot-attacks-on-healthcare-data/
[15] "Cybercrime Trends", Website: https://risk.lexisnexis.com/insights-resources/research/cybercrime-report
[16] "AN INTELLIGENT WEB-BASED VOICE CHAT BOT" by:
[17] S. J. du Preez, Student Member, IEEE, M. Lall, S. Sinha, MIEEE, MSAIEE
[18] "Humans and Bots in Internet: Measurement, Analysis, and Automated Classification" by:
[19] Steven Gianvecchio, MengjunXie, Member, IEEE, Zhenyu Wu, and Haining Wang, Senior Member, IEEE
[20] "Bot and Intelligent Agent Research Resources 2020" by:
[21] Marcus P. Zillman, M.S., A.M.H.A., Executive Director – Virtual Private Library