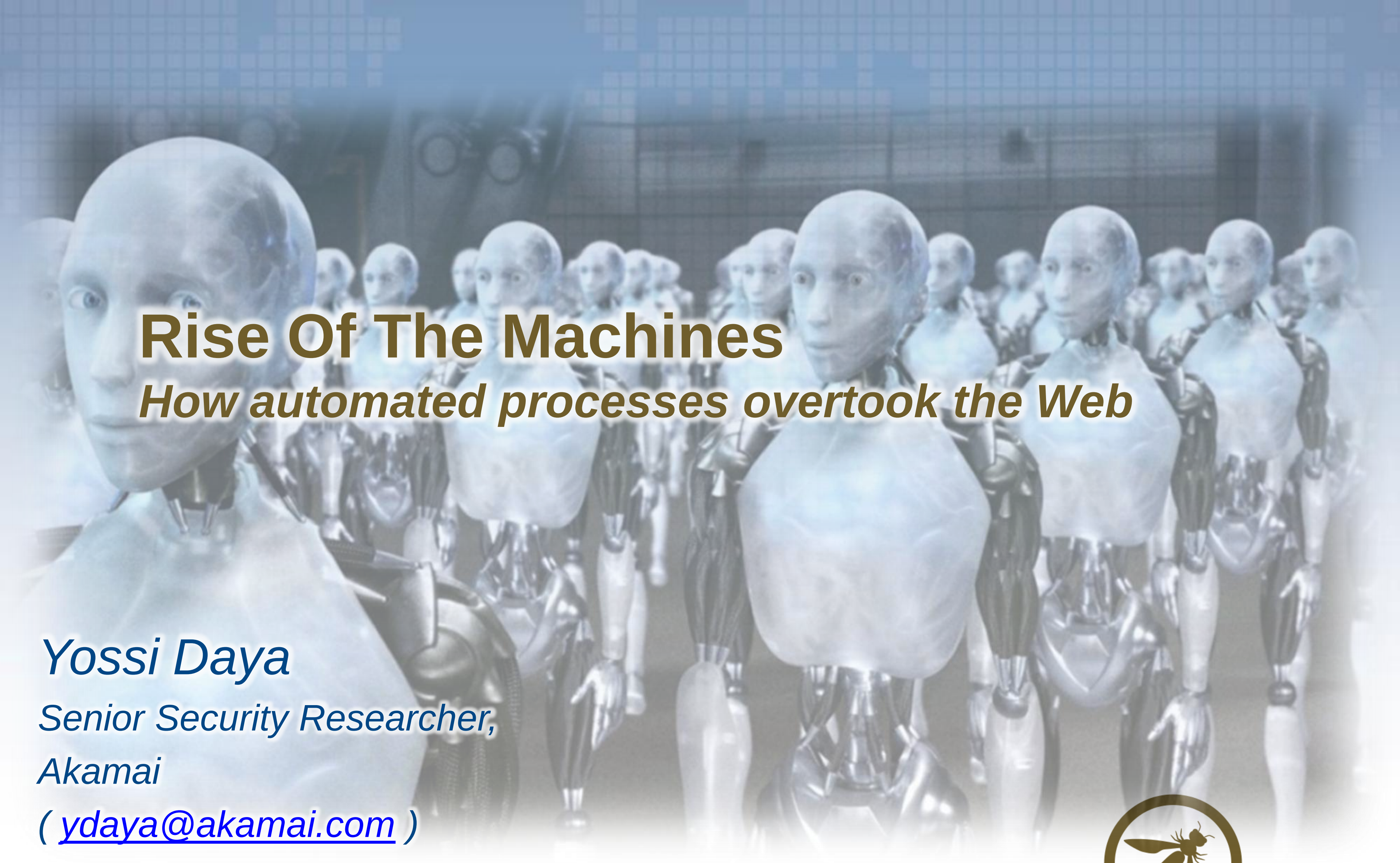


Rise Of The Machines

How automated processes overtook the Web

Yossi Daya

Senior Security Researcher,

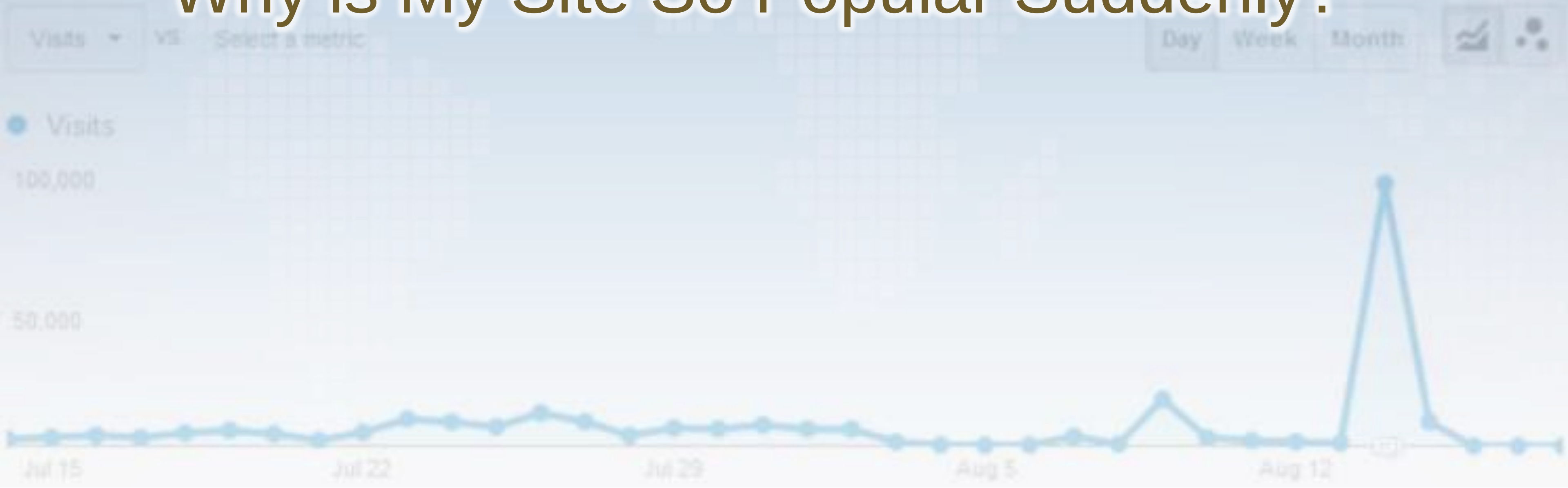
Akamai

(ydaya@akamai.com)

OWASP AppSecEU 15

Amsterdam, The Netherlands

Why is My Site So Popular Suddenly?



Real World Customer Story

- Large retail site detects traffic increase
  - Increase is not related with a small set of IPs:
    - ✗ DDoS ruled out
    - ✗ Scraper ruled out
    - ✗ No attack traffic – web attacks ruled out
    - ✓ Thousands of new IPs
    - ✓ Each IP browses for products
    - ✓ Each IP creates small number of requests

It feels like a DDoS – but it isn’t…

It feels like a scraper, but it isn’t…

It’s not a web attack… What is it?

Big Data Analysis – Behavioral Profiling

- 3000 IP's from 15 different subnets
- Each IP requested <100 requests
- All IPs belong to the same cloud provider
- All IPs requested the same folder but changed the "file name" (or product ID)
  - /Product/XXXXX

Big Data Analysis – Signatures

- Single User-Agent

Mozilla/5.0 (Windows NT 6.1; WOW64; rv:29.0) Gecko/20100101 Firefox/29.0

- Common HTTP request headers

Connection : close

X-Forwarded-For : unknown

And the award goes to….

Highly distributed “Mega-Scraper”

- This mega-scraper was generating millions of HTTP requests – mainly product searches
- Bot Net scraped 7 different large retail websites using the same method
- Only by looking at this as a “wide phenomena” you get the true nature of the beast

OWASP AppSecEU 15
Amsterdam, The Netherlands

Security Big Data at Akamai:
Cloud Security Intelligence

20 Terabytes of daily attack data

2 Petabytes of security data stored

Up to 90 days retention

600K log lines/sec. indexed by 30 dimensions

8000 queries daily scanning terabytes of data

24hrs.

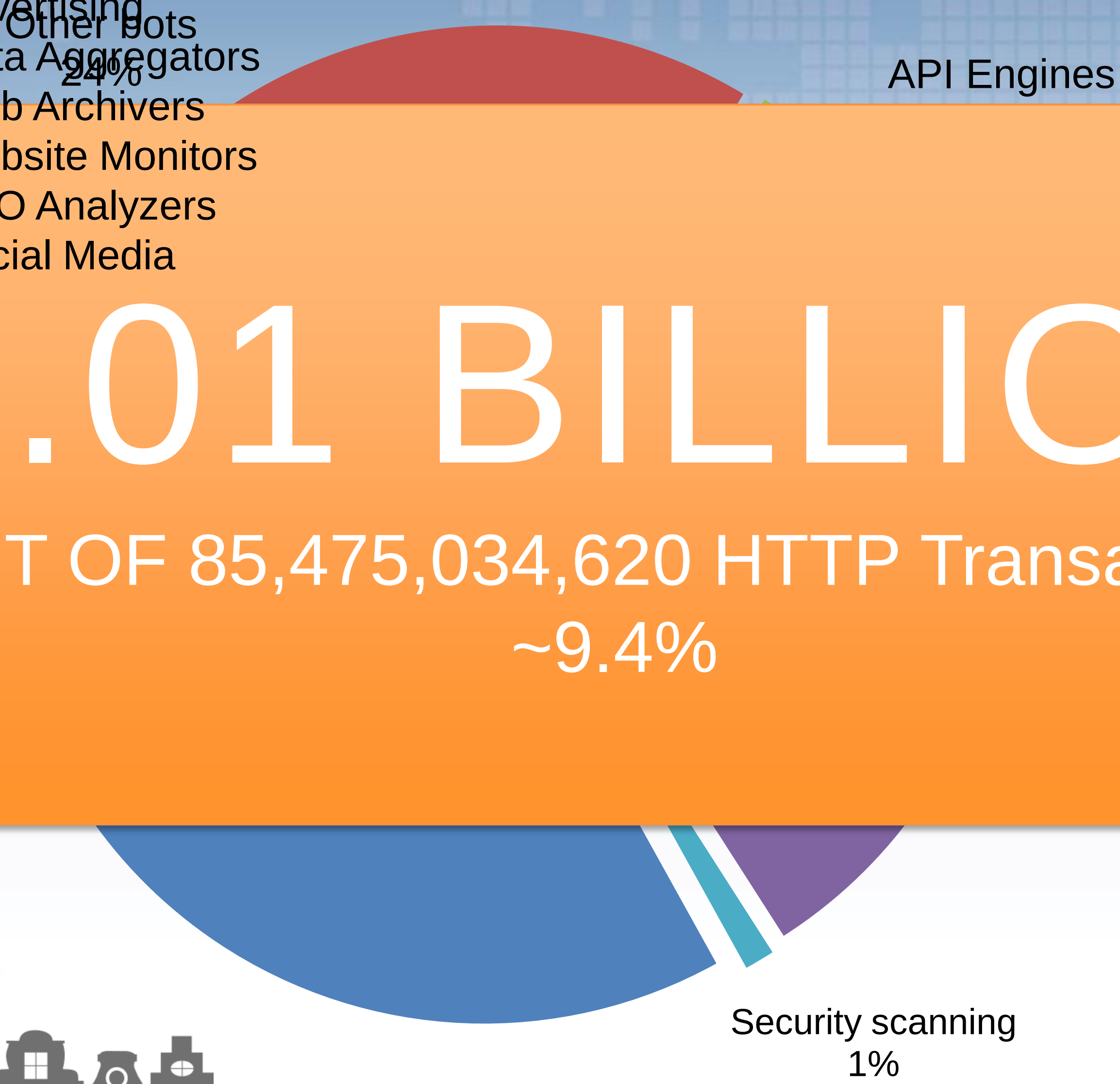
How many web bots do we see in one day?

- 24% Content Scrapers
- 7% Advertising
- 3% Data Aggregators
- 2% Web Archivers
- 2% Website Monitors
- 1% SEO Analyzers
- 1% Social Media

8.01 BILLION

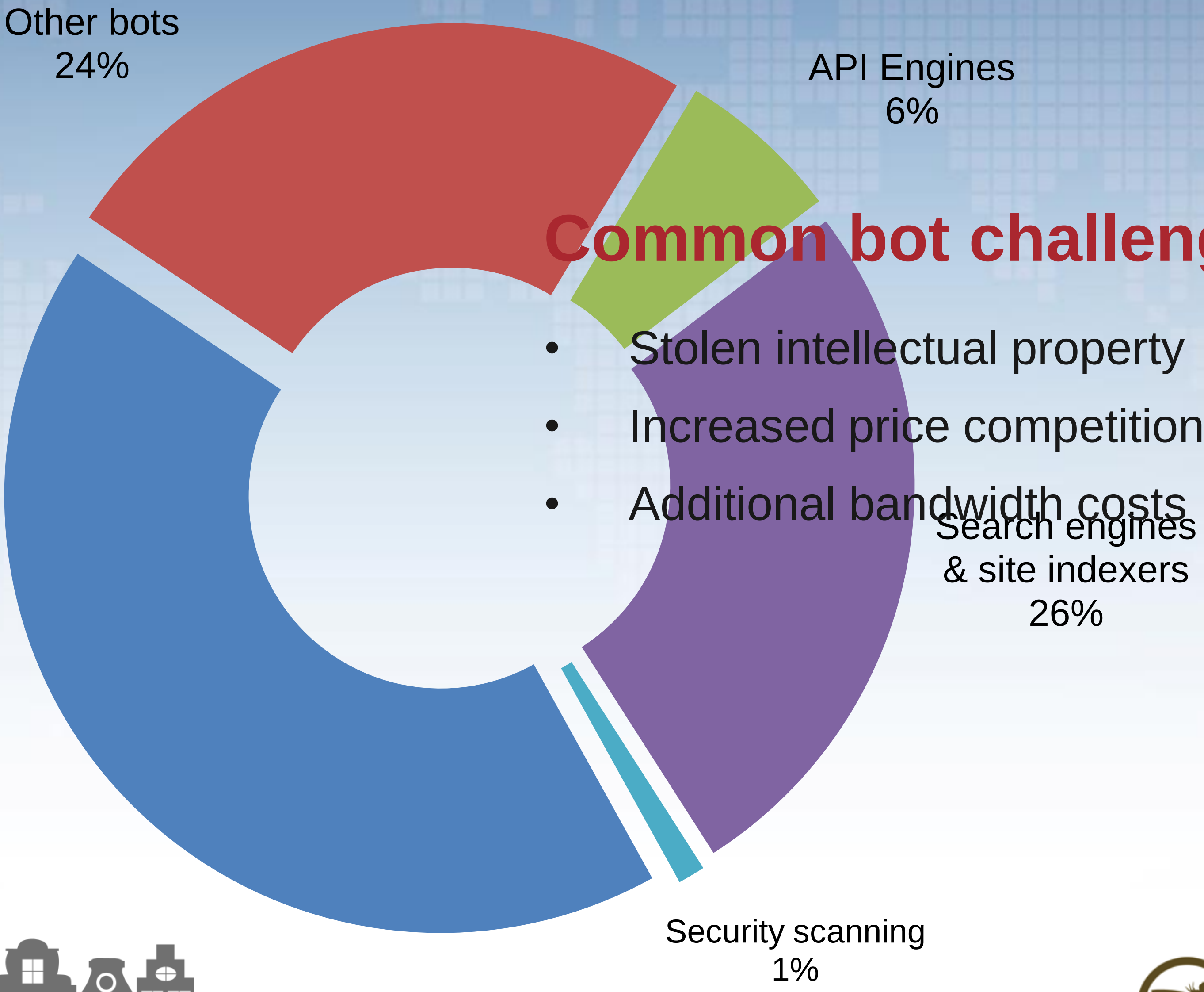
OUT OF 85,475,034,620 HTTP Transactions

~9.4%



Common bot challenges

- Stolen intellectual property
- Increased price competition
- Additional bandwidth costs



OWASP AppSecEU 15
Amsterdam, The Netherlands

DETECTING BOTS…

Detection Methods

- Transactional Based
  - Signatures
  - HTTP quirks
- Behavioral
  - Big data analytics
    - Observation over time
- Rate Controls
- Human vs. Bot challenge

Signatures

Who are you ?

- Declared bots :
  - User Agent Identification (name, description, URL, Email)
  - HTTP request header identification (“From:”)

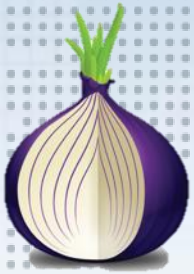
MyBot/1.0 (+http://mysite.com/mybot)

What platform are you using ?

- Detected bots :
  - User Agent detection
  - HTTP request header detection (header ordering)
- Development platforms
- Http Libraries
- Scraping platforms (libraries, services)
- Headless browsers/Automation tools

Where are you coming from ?

IP source is a good indication…


Tor

Cloud Infrastructure

Quirks

- User Agent quirks

```
User-Agent: User-Agent: Mozilla/5.0 (Windows NT 6.1; WOW64)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/36.0.1985.125
Safari/537.36

User-Agent: ‘Mozilla/5.0 (Windows; U; Windows NT 6.1; en-US)
AppleWebKit/534.16 (KHTML, like Gecko) Chrome/36.0.1985.125
Safari/534.16’

User-Agent: Mozilla/5.0 (Windows; U; Windows NT 6.1; en-US)
AppleWebKit/534.16 (KHTML, like Gecko) Chrome/50.0.648.204
Safari/534.16
```

Is there something weird ?

HTTP request headers quirks…

- Small numbers of headers (only host, connection and user agent)
- `accpet : *.*`
- Duplicate header names
- HTTP/1.0 and lower
- “Connection : close”

OWASP AppSecEU 15
Amsterdam, The Netherlands

Are you really you ?

Search engine impersonators

Looks like known search engine but

Originates from different networks

Behavioral Profiling

OWASP AppSecEU 15
Amsterdam, The Netherlands

Activity overall

- Big data analysis - 6-12 hours traffic
  - How long a single IP been active on site ?
  - How many different resources were requested ?
    - Same page, multiple queries
    - Same host, multiple paths
  - Are there regular patterns over time ?

Target Resources

- Same page, multiple queries
  - Is he looping values for query parameter? (?product_id = XXXX)
- Same host, multiple paths
  - Is he looping through path “file names”

    (/Product/XXXX)

Website response code ratio

302 Found

200 OK

404 Not Found

401 Unauthorized

403 Forbidden

OWASP AppSecEU 15
Amsterdam, The Netherlands

Workflow

- Does the IP follow a legitimate user workflow ?
  - Homepage → Search page
  - Add product → Shopping cart
  - Search page → Autocomplete page

OWASP AppSecEU 15
Amsterdam, The Netherlands

Bot Net

- Distributed Bot using multiple IP’s :
  - One or more network (subnets, AsNumbers)
  - Same set of User Agents
  - Common HTTP request header signature
  - Requesting same resources

Mitigation & Management

OWASP AppSecEU 15
Amsterdam, The Netherlands

Should we stop bots ?

- Not a security problem
- Not necessarily bad
  - Search engines
  - Price comparisons
- They will always come back
  - More sophisticated
  - Harder to detect

OWASP AppSecEU 15
Amsterdam, The Netherlands

Management

- Managing bots
  - Approve full access
  - Slow them down
  - Serve stale objects
  - Activity time limit

Summary

- Large portions of web site traffic is generated by automated bots
- While signature-based detection will go a long way, big data analytics is required in order to detect distributed activities which are the de-facto method today
- While it’s not a security problem per-se, businesses lose revenue
- Attempting to stop bots will only make things worse

Thank You

OWASP AppSecEU 15
Amsterdam, The Netherlands