Behind the Scenes of Online Attacks: an Analysis of Exploitation Behaviors on the Web

Davide Canali, Davide Balzarotti

Software and System Security Group
EURECOM, France
http://s3.eurecom.fr/
Motivations

- Studying the internals of web attacks
  - What attackers do while and after they exploit a vulnerability on a website
  - Understand why attacks are carried out (fun, profit, damaging others, etc.)

- Previous studies
  - how attacks against web sites are carried out
  - how criminals find their victims on the Internet
  - Lack of studies on the behavior of attackers (what they do during and after a typical attack)
    » Previous works used static, non functional honeypots (not exploitable)
How

- 2500 vulnerable applications deployed on 500 websites on 100 domain hosted on 9 popular hosting providers
  - 5 common CMSs (blog, forum, e-commerce web app, generic portal, SQL manager), 1 static website and 17 PHP web shells
Data collection

- **100 days** of operation
- **Centralized data collection** for simple and effective management
- Each deployed website acts as a proxy
  - Redirects traffic to the **real web applications** installed on **VMs** in our premises
Collected data

- ~10 GB of raw HTTP requests
- In average:
  - 1-10K uploaded files every day
  - 100-200K HTTP requests/day
- First suspicious activities:
  - automated: 2h 10' after deployment
  - manual: after 4h 30'

Requests volume
Requests by country
(excluding known crawlers)

- Color intensity is logarithmic!
- IPs from the USA, Russia and Ukraine account for 65% of the total requests
1. **Discovery**: how attackers find their targets
   - Referer analysis, dorks used to reach our websites, first suspicious activities

69.8% of the attacks start with a scout bot visiting the pages often disguising its User-Agent.
Attack analysis
The four different phases

1. **Discovery**: how attackers find their targets
   - Referer analysis, dorks used to reach our websites, first suspicious activities

2. **Reconnaissance**: how pages were visited
   - Automated systems and crawling patterns identification, User-Agent analysis

69.8% of the attacks start with a scout bot visiting the pages often disguising its User-Agent

In 84% of the cases, the attack is launched by a 2nd automated system, not disguising its User-Agent (exploitation bot)
1. **Discovery**: how attackers find their targets
   - Referer analysis, dorks used to reach our websites, first suspicious activities

2. **Reconnaissance**: how pages were visited
   - Automated systems and crawling patterns identification, User-Agent analysis

3. **Exploitation**: attack against the vulnerable web app
   - Exploits detection and analysis, exploitation sessions, uploaded files categorization, and attack time/location normalization
   - Analysis of forum activities: registrations, posts and URLs, geolocation, message categories

---

- 69.8% of the attacks start with a scout bot visiting the pages often disguising its User-Agent
- In 84% of the cases, the attack is launched by a 2nd automated system, not disguising its User-Agent (exploitation bot)
- 46% of the successful exploits upload a web shell
1. **Discovery**: how attackers find their targets
   - Referer analysis, dorks used to reach our websites, first suspicious activities

2. **Reconnaissance**: how pages were visited
   - Automated systems and crawling patterns identification, User-Agent analysis

3. **Exploitation**: attack against the vulnerable web app
   - Exploits detection and analysis, exploitation sessions, uploaded files categorization, and attack time/location normalization
   - Analysis of forum activities: registrations, posts and URLs, geolocation, message categories

4. **Post-Exploitation**: second stage of the attack, usually carried out manually (optional)
   - Session identification, analysis of shell commands

---

**Attack analysis**

The four different phases

- **69.8% of the attacks start with a scout bot visiting the pages often disguising its User-Agent**
- **In 84% of the cases, the attack is launched by a 2nd automated system, not disguising its User-Agent (exploitation bot)**
- **46% of the successful exploits upload a web shell**
- **3.5 hours after a successful exploit, the typical attacker reaches the uploaded shell and performs a second attack stage for an average duration of 5' 37”**
Attack analysis
phases #1-2: discovery - reconnaissance

- **Discovery: referer** shows where visitors are coming from
  - Set in 50% of the cases
  - Attackers find our honeypots mostly from search engine queries (in the order: Google, Yandex, Bing, Yahoo)
  - Some visits from web mail services (spam or phishing victims) and social networks
Attack analysis
phases #1-2: discovery - reconnaissance

• **Discovery:** referer shows where visitors are coming from
  - Set in 50% of the cases
  - Attackers find our honeypots mostly from search engine queries (in the order: Google, Yandex, Bing, Yahoo)
  - Some visits from web mail services (spam or phishing victims) and social networks

• **Reconnaissance:** how were pages visited?
  - 84% of the malicious traffic was from automated systems
    » No images or style-sheets requested
    » Low inter-arrival time
    » Multiple subdomains visited within a short time frame
  - 6.8% of the requests mimicked the User-Agent string of known search engines
Attack analysis
phase #3: exploitation

- **444 distinct exploitation sessions**
  - Session = a set of requests that can be linked to the same origin, arriving within 5' from each other
  - 75% of the sessions used at least once 'libwww/perl' as User-Agent string → scout bots and **automatic attacks**

- **Almost one exploitation out of two uploaded a web shell**, to continue the attack at a later stage (post-exploitation)
**Attack analysis**

phase #3: Forum activity

- Daily averages: 604 posts, 1907 registrations, 232 online users
  - One third of the IPs acting on the forum registered at least one account, but never posted any message → any business related to selling forum accounts?

- ∼1% of the links posted to the forum led to malicious content†

† According to Google SafeBrowsing and Wepawet
Attack analysis
phases #3-4

- Clear hourly trends for post-exploitation (manual) sessions
Attack analysis
phase #4: post-exploitation

- Almost **8500 interactive sessions** collected
  - Known and unknown web shells
  - **Average session duration: 5' 37”**
    » 9 sessions lasting more than one hour
  - **Parsed commands from the logs**
    » 61% of the sessions **upload a file** to the system
    » 50% of the sessions (try to) **modify existing files**
      • Defacement in 13% of the cases
Attacker goals

- The analysis of collected files allows to understand the attackers' goals
  - File normalization and similarity-based clustering
  - Manual labeling of clusters
Clustering example

- Similarity clustering on web shells (ours are labeled)
Conclusions

• The study confirmed some known trends
  – Strong presence of Eastern European countries in spamming activities
  – Scam and phishing campaigns often run from African countries
  – Most common spam topic: pharmaceutical ads

• Unexpected results
  – Most of the attacks involve some manual activity
  – Many IRC botnets still around
  – Despite their low sophistication, these represent a large fraction of the attacks to which vulnerable websites are exposed every day
Thank you

? 

For further questions, suggestions, comments: canali@s3.eurecom.fr

Special thanks to Master students helping me with:

- Log analysis
  » Marco Pappalardo
  » Roberto Jordaney

- File analysis
  » Maurizio Abbà