Monkey-Spider

Detection of Malicious Web Sites

Final presentation of the diploma thesis

Ali Ikinci
ali[at]ikinci.info

9. July 2007

Head of Department: Prof. Dr. Felix Freiling
Supervisor: Dipl.-Inform. Thorsten Holz
Laboratory for Dependable Distributed Systems
UNIVERSITY OF MANNHEIM

Outline

- Problem and challenge
- Simplified architecture
- Requirements analysis
- Honeypots vs. honeyclients
- Monkey-Spider architecture
- Limitations
- Preliminary results
- Key Findings

Problem

- Client side attacks are on the rise
- Many abuses of the Internet[1][2][3]
- No comprehensive and free database of threats on the Internet
- HoneyMonkey[4]

STRIDER

- SiteAdvisor[5]

McAfee SiteAdvisor

A sample SiteAdvisor site report

HOME | DOWNLOAD | ANALYSIS | SUPPORT | BLOG | ABOUT US | Look up a site report:

zango.com

In our tests, we found downloads on this site that some people consider adware, spyware, or other unwanted programs.

Are you the owner of this site? Leave a comment

Contact information:	Established in	Country	Popularity
	1998	United States	Lots of users

AUTOMATED WEB SAFETY TESTING RESULTS FOR ZANGO.COM

E-MAIL TESTS FOR ZANGO.COM:

< 1 e-mail/ month

After entering our e-mail address on this site, we received less than 1 e-mail per month.

View detailed analysis

What our inbox looked like after we signed up here:

Sample Inbox

Subject	Sender	Date
Message From: SiteAdvisor Assistant	Unknown Sender	June 2006

DOWNLOAD TESTS FOR ZANGO.COM:

52 red downloads

In our tests, we found downloads on this site that some people consider adware, spyware, or other unwanted programs.

View detailed analysis

Submit a download for analysis

See how McAfee can protect your PC from dangerous downloads.

Downloads we found on this site:

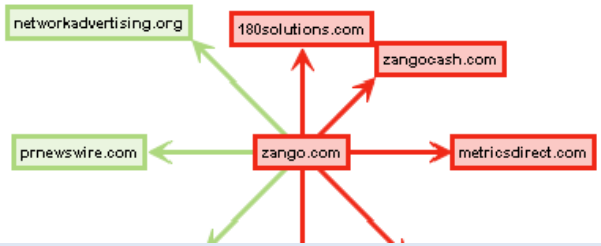
Download	Analysis
Add another account? (GauntletMessengi	Adware-180solutions
Bidulator.exe	Adware-180solutions
Checkers (ZangoCheckers.exe)	Adware-180solutions
Chess (ZangoChess.exe)	Adware-180solutions
David vs Goliath 1.0.0.0 (setupdavid2421.	Adware-180solutions

52 total downloads. See more.

ONLINE AFFILIATIONS FOR ZANGO.COM:

Links to red sites

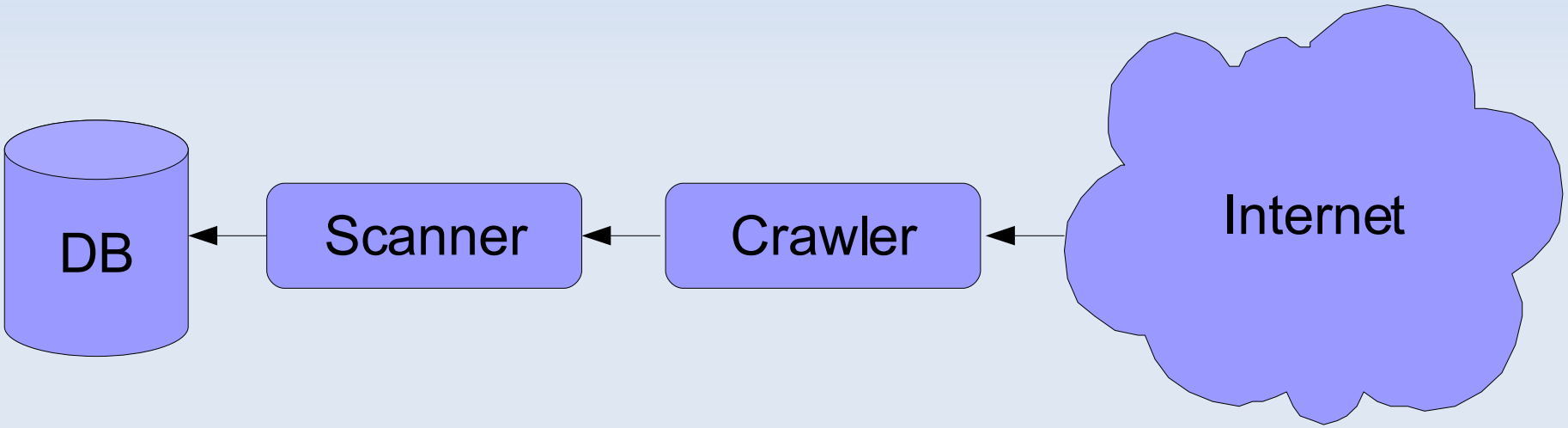
When we tested this site we found links to 180solutions.com, which we found to be a distributor of downloads some people consider adware, spyware or other unwanted programs.



Challenge

- Find actual threats and zero-day exploits on the Internet
- Collect malicious code
- Allow various infection vectors
- Build a database with detailed relevant information about threats
- Continuous monitoring of suspicious resources

Simplified Architecture of the Monkey-Spider system



Requirements Analysis

- Performance
- Modularity and Expandability
- Multithreaded modules
- Parallel operation
- Scalability
- Usability

Requirements Analysis

- Crawler part:
  - Crawling policies
  - Link extraction
  - URL normalization
  - Efficient storage

Requirements Analysis

- Malware scanner:
  - Multiple malware scanners
  - Support for automated dynamic malware analysis tools
  - Expandability
- Database
  - Store relevant information
  - Bunch of standard querys

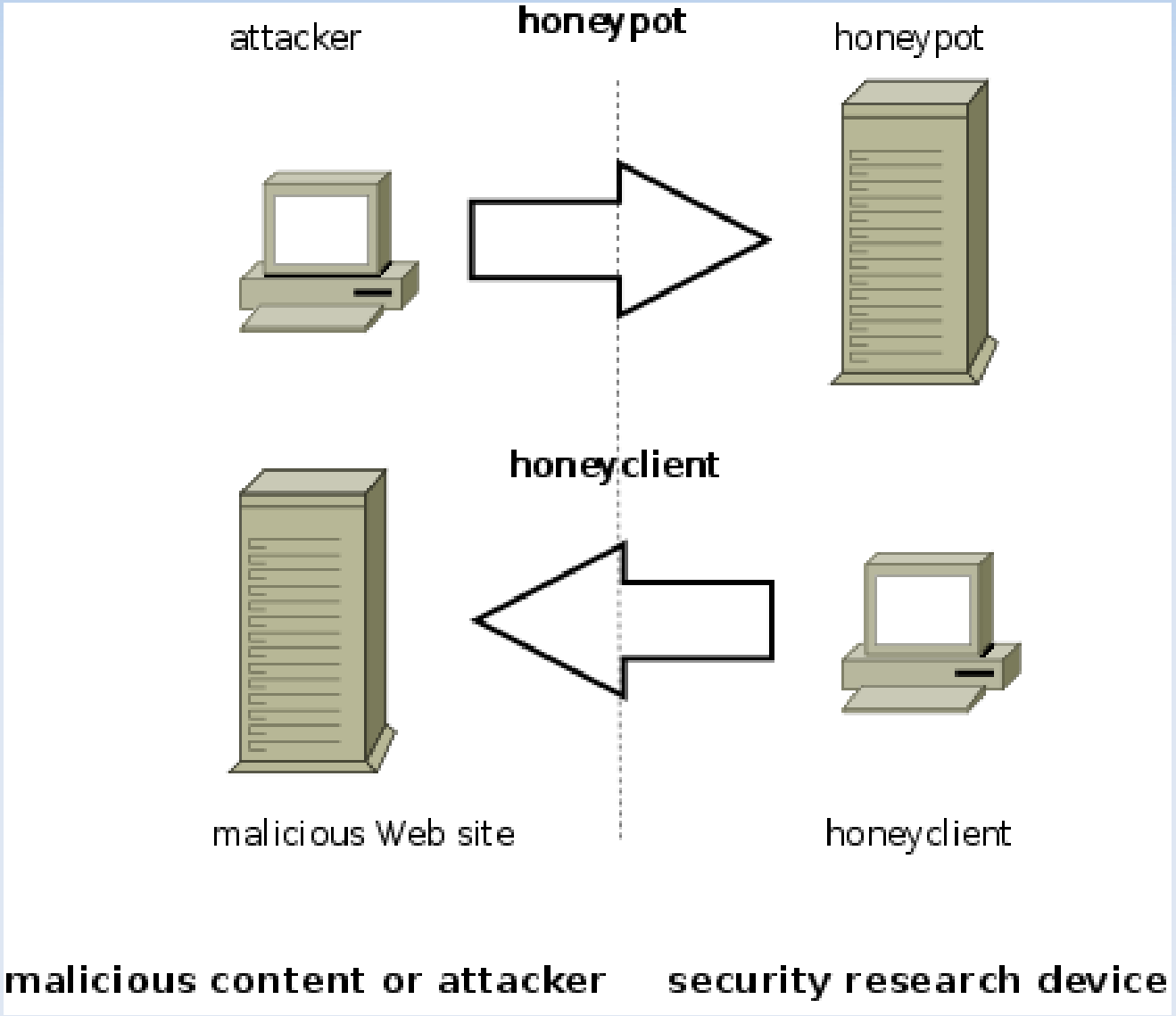
Solution Ideas

- Do not reeinvent the wheel
- Use existing Free Software
- Use existing honeypot technologies
- Use extensive prototyping

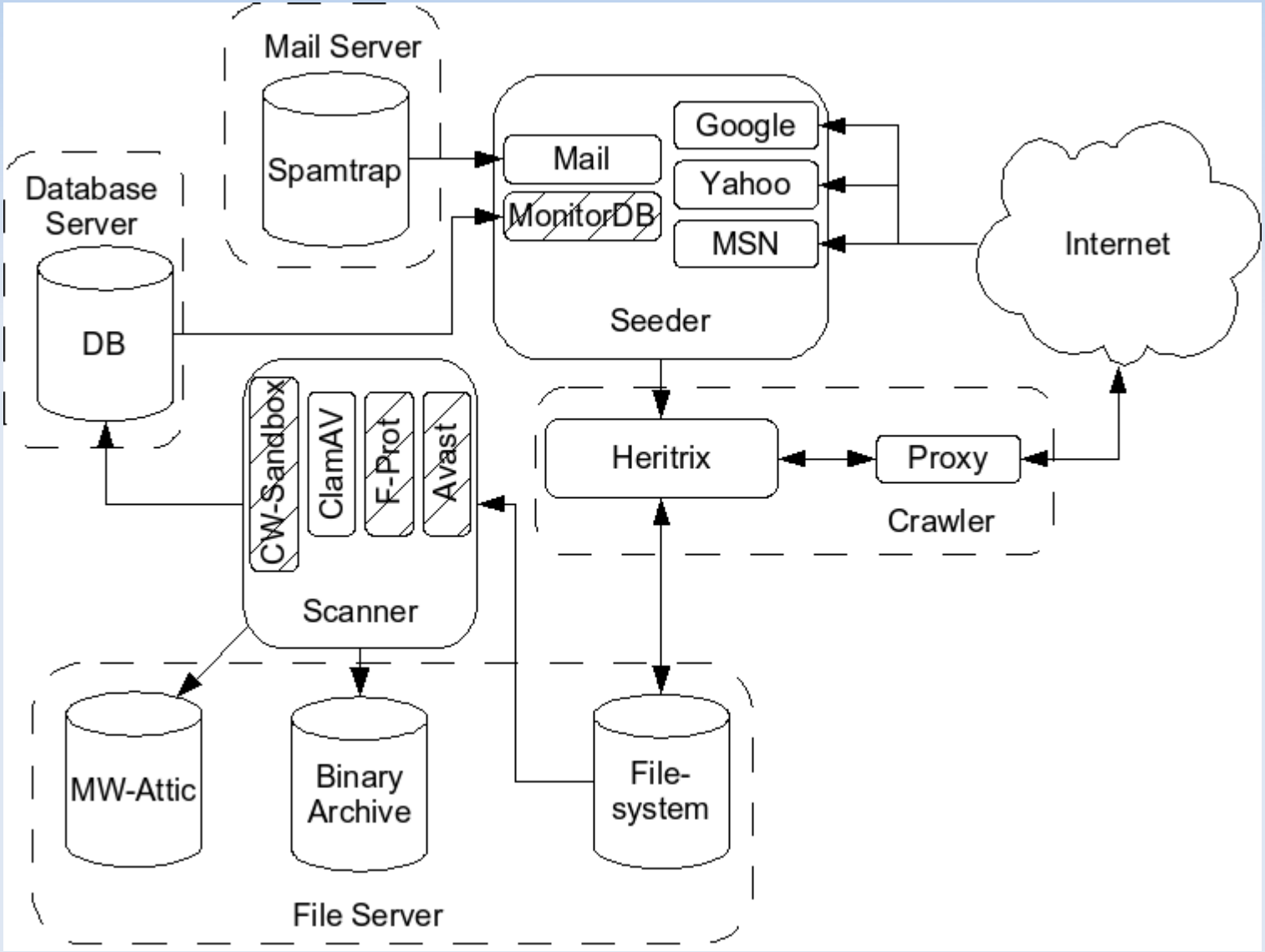
Honeypots

- Honeypots are dedicated deception devices
- Two types:
  - server honeypots or honeypots and
  - client honeypots or honeyclients
- Both can be classified as:
  - low-interaction honeypots or
  - high-interaction honeypots
- Similar Web maliciousness detection systems operate either as low- or high-interaction honeyclients
- The Monkey-Spider system operates as a crawler based low-interaction honeyclient

Honeypot vs. Honeyclient



Monkey-Spider: Architecture



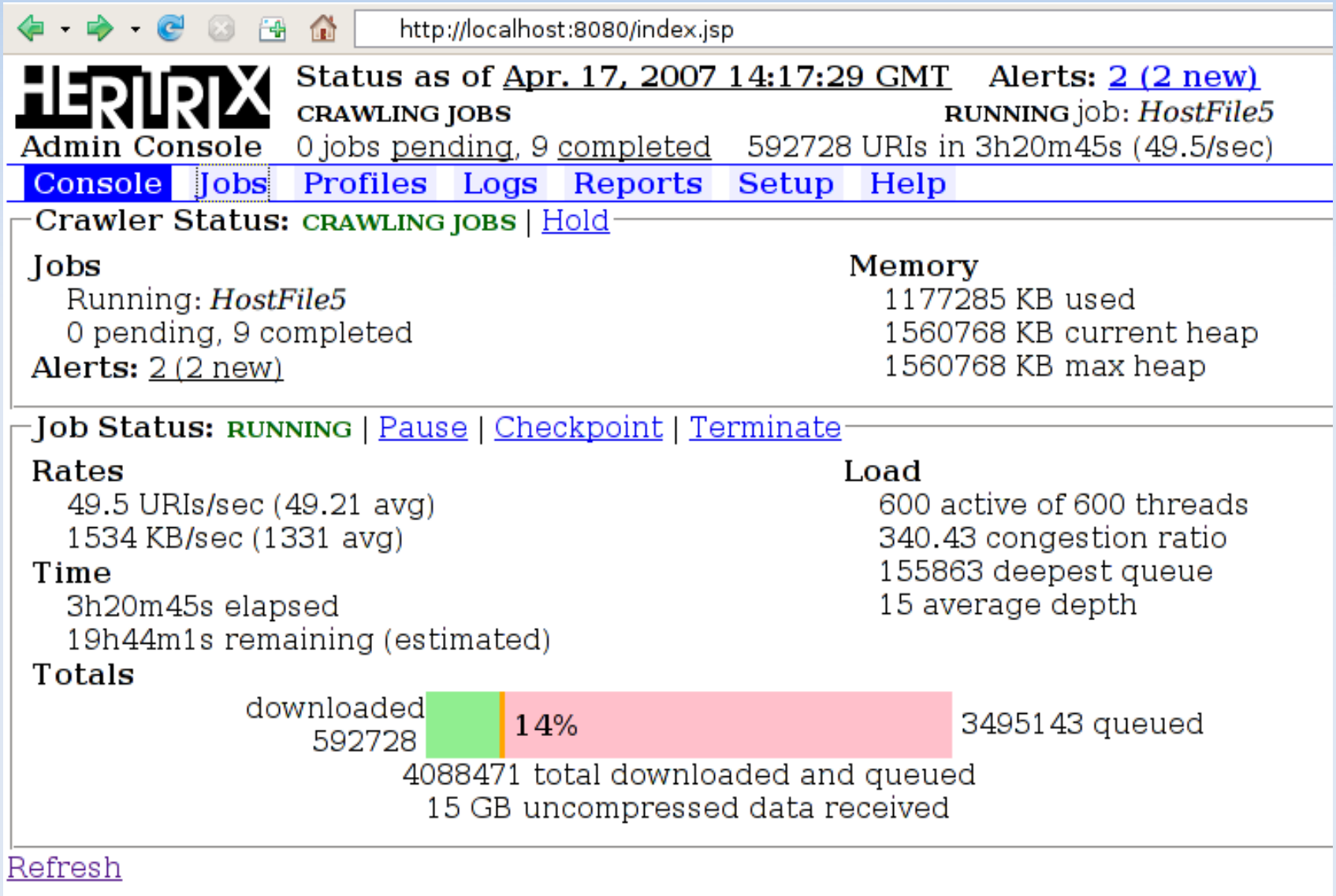
Monkey-Spider: Queue Generation

- Provide starting point(s) (seeds) utilizing different approches:
  - Web search seeders (Google, MSN and Yahoo)
  - (Spam) mail seeder
  - Hosts file seeder
  - Monitoring seeder

Heritrix WebCrawler[6]

- Built for the Internet Archive
- Free Software
- Recursive, scalable and multithreaded crawling
- Thouroughly tested
- Continously extended
- Many parameters
- Controled with
  - Web interface
  - Java Management Extensions (JMX)
- Generates ARC-files as output

The Heritrix Web Interface



ARC File-Format

- Designed by the Internet Archive
- Large aggregate files for ease of storage
- Features:
  - self-contained
  - multi-protocol able
  - streamable
  - viable

Sample:

```
http://www.dryswamp.edu:80/index.html\
127.10.100.2 19961104142103 text/html
202
HTTP/1.0 200 Document follows
Date: Mon, 04 Nov 1996 14:21:06 GMT
Server: NCSA/1.4.1
Content-type: text/html Last-modified:\
Sat,10 Aug 1996 22:33:11 GMT
Content-length: 30
<HTML>
Hello World!!!
</HTML>
```

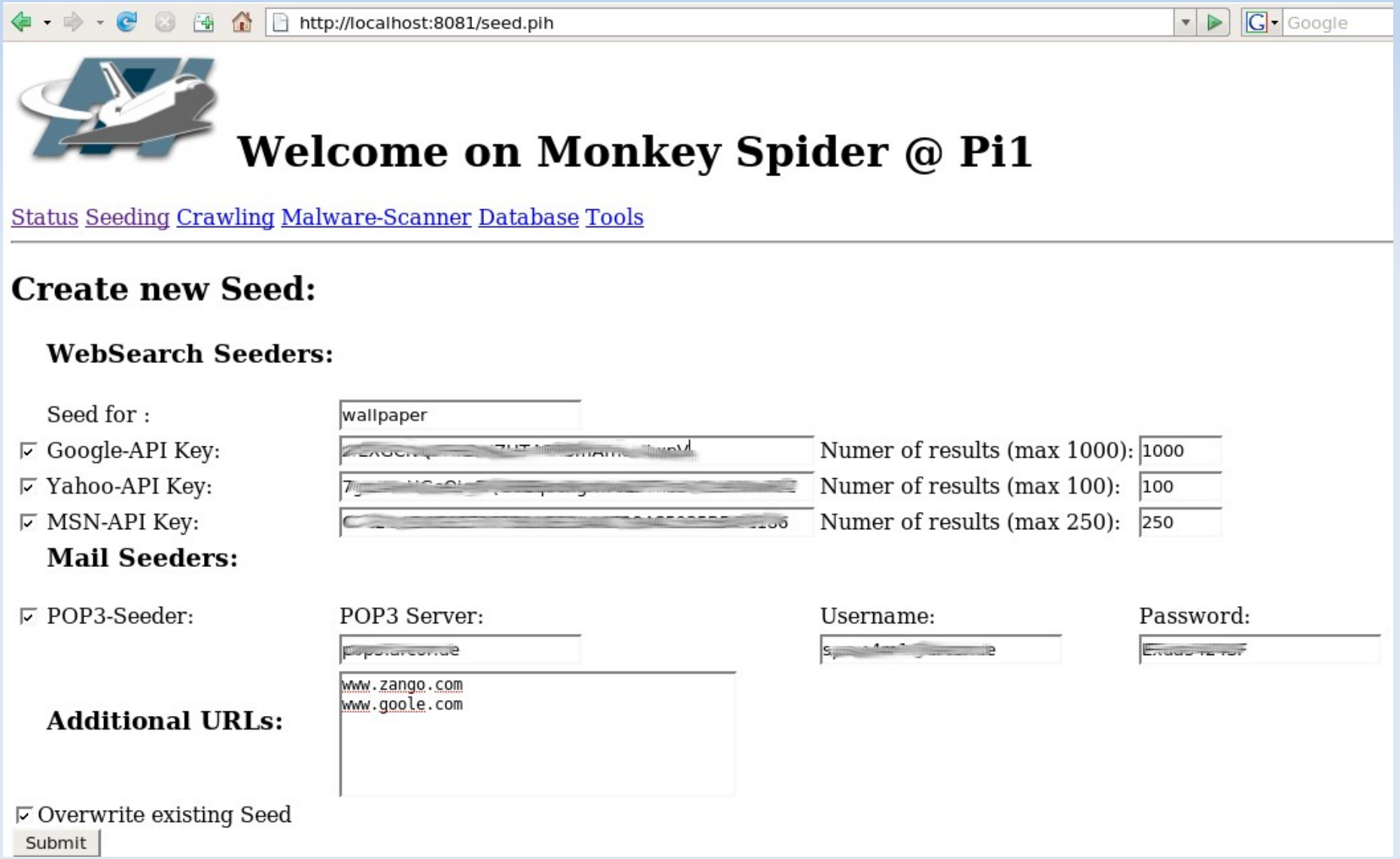
Malware Scanner

- ARC-Files are unpacked and examined
- MW-Scanners are executed on crawled content
  - Found Malware is stored
- Information regarding the malware is stored into database

The Monkey-Spider Web interface

- Controles the whole system
- Modules are seperately manageable
- Standard querys are provided
- Job based
- Authentification

The Seed generation page



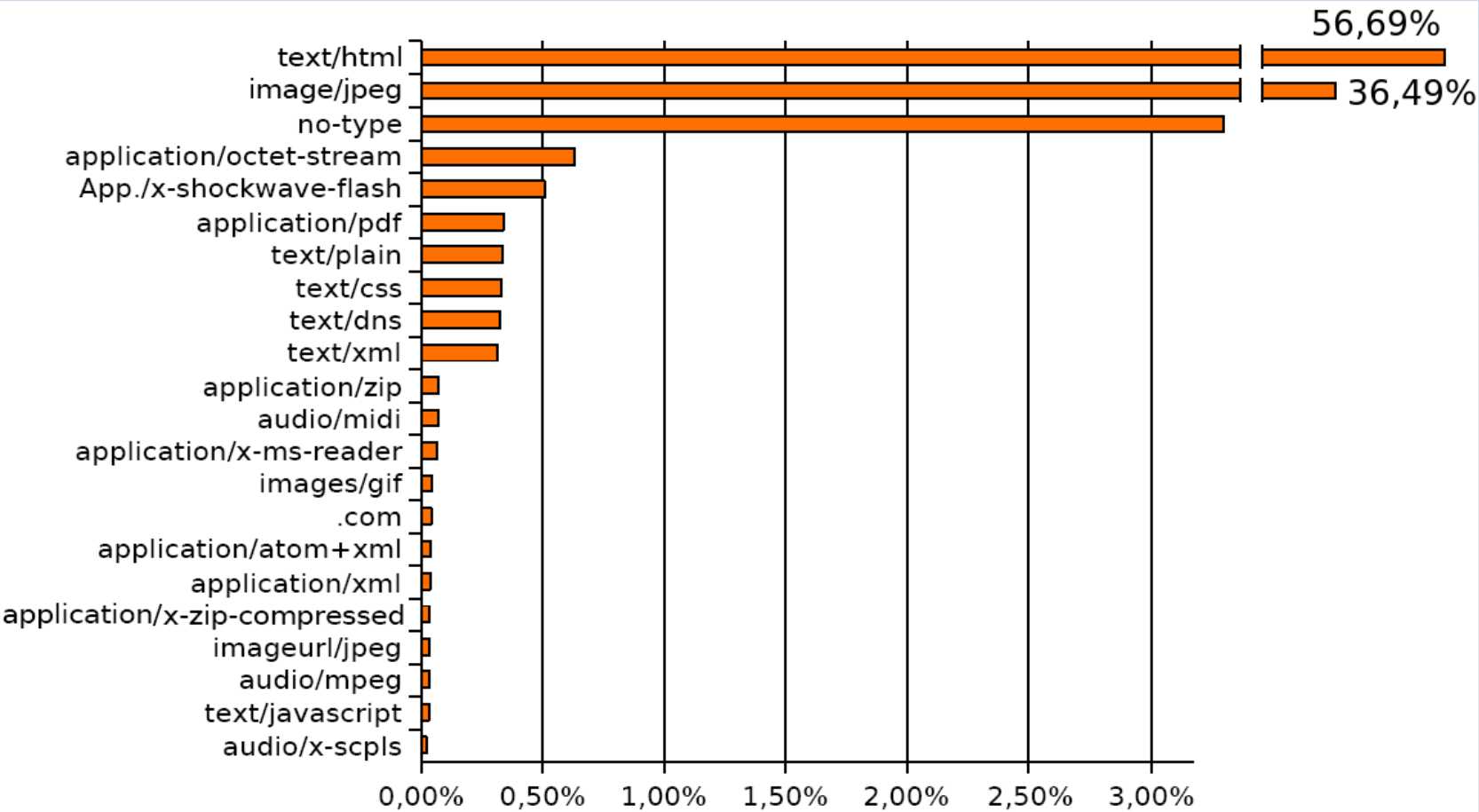
Limitations

- Analysis is limited to the publicly indexable web[7]
- Only known malware is recognized and stored
  - Will be enhanced with CWSandbox
- Drive-by download sites, heavily obfuscated JavaScript code and zero-day exploits are not recognized
- Full scan of the Web is not possible with Heritrix yet
- Two seperate jobs are not aware of examining the same sites and contents

Preliminary Results

- We have done various crawls over two months
- We crawled for various topics and did a hosts file based crawl
- defective crawl settings caused incomplete preliminary results

MIME-type distribution of crawled content



Topic based maliciousness

topic	maliciousness in %
pirate	2.6
wallpaper	2.5
hosts file	1.7
games	0.3
celebrity	0.3
adult	0.1
total	1

Top 10 malware sites

domain	occurence
desktopwallpaperfree.com	487
waterfallscenes.com	92
pro.webmaster.free.fr	91
astalavista.com	15
bunnezone.com	14
oss.sgi.com*	12
ppd-files.download.com	12
888casino.com	11
888.com	11
bigbenbingo.com	10

* non malicious Web site (false positive)

Top-10 malware types

name	occurence
HTML.MediaTickets.A	487
Trojan.Aavirus-1	92
Trojan.JS.RJump	91
Adware.Casino-3	22
Adware.Trymedia-2	12
Adware.Casino	10
Worm.Mytob.FN	9
Dialer-715	8
Adware.Casino-5	7
Trojan.Hotkey	6

Key Findings

- 1% of all examined Web sites are malicious
- adult Web sites are relative harmless
- most malware is spread through pirate and wallpaper propagation Web sites
- to gather representative results a Web site has to be completely crawled and analysed
- the scope of the crawl has to be choosen carefully
- We know very little about malicious Web sites and their operators

Performance

- We measured the performance of our crawls on a standard PC
- Crawl performance of 1 MB/sec
- Malware analysis (without the crawling) in 0.05 seconds per downloaded content and 2.35 seconds per downloaded and compressed MB
- Resulting in about 3.35 seconds per analysed MB of content
- In comparison: other low-interaction honeyclient based Web analysers require a minimum of 3 seconds per Web site

Future Trends

- Attacks are concentrated more and more from the server to the client
- Client programs other than the Web client are targeted more often, like Media Players, Flash and PDF interpreters
- Advanced honeypot, virtual machine and anti-virus program detection techniques contained in malware complicates the detection of such

Live - Demo

- Live demonstration of the current state of Monkey-Spider

Questions ?

Thank you for your attention!

References

[1] Anti-Phishing Working Group (APWG) „Phishing Activity Trends Report, Combined Report for September and October“ 2006 http://www.antiphishing.org

[2] Thorsten Holz, „A Short Visit to the Bot Zoo“, IEEE Security & Privacy , 2005, volume 3, number 3, pages 76-79

[3] S. Saroiu, S. D. Gribble, and H. M. Levy „Measurement and Analysis of Spyware in a University Environment“ USENIX Proceedings of the 1st Symposium on Networked Systems Design and Implementation (NSDI), San Francisco, CA, March 2004

[4] The Strider HoneyMonkey Project http://research.microsoft.com/HoneyMonkey/

[5] McAfee SiteAdvisor http://www.siteadvisor.com/

[6] Heritrix the Internet Archive's WebCrawler http://crawler.archive.org/

[7] Lawrence, S. and Giles, C. L. 2000. Accessibility of information on the Web. Intelligence 11, 1 (Apr. 2000), 32-39.