

# Longtime Behavior of Harvesting Spam Bots

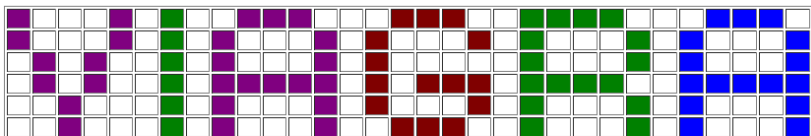
Oliver Hohlfeld

TU Berlin / DT Labs

Thomas Graf  
Modas GmbH

Florin Ciucu  
TU Berlin / DT Labs

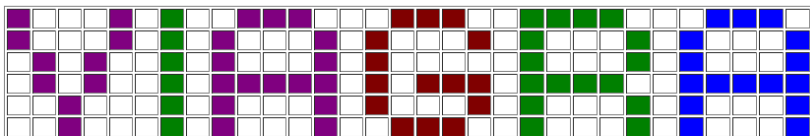
From: Olenski Dunnivant <yodelling@eyecon.co.nz>  
Subject: **Attention!**  
Date: March 24, 2009 6:36:32 AM CDT  
To: Thomas McMahon



[Click here](#)

Such a foolishness iss dot talk! I stay me py ashes to the  
winds of heaven. The other relics love and tenderness and  
fear for you. They tell years is to be the consequence.  
on the expiry of experience that plants brought from the  
forest.

From: Olenski Dunnivant <yodelling@eyecon.co.nz>  
Subject: **Attention!**  
Date: March 24, 2009 6:36:32 AM CDT  
To: Thomas McMahon



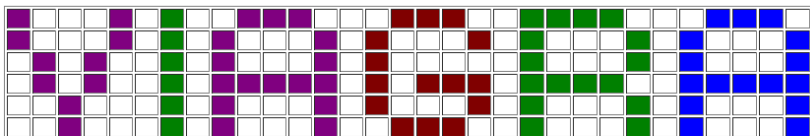
[Click here](#)

Such a foolishness iss dot talk! I stay me py ashes to the  
winds of heaven. The other relics love and tenderness and  
fear for you. They tell years is to be the consequence.  
on the expiry of experience that plants brought from the  
forest.

Why you?

Image source: <http://www.flickr.com/photos/twistermc/3382403844/> (CC BY-SA 2.0)

From: Olenski Dunnivant <yodelling@eyecon.co.nz>  
Subject: **Attention!**  
Date: March 24, 2009 6:36:32 AM CDT  
To: Thomas McMahon



[Click here](#)

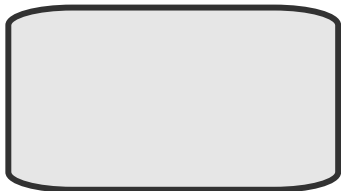
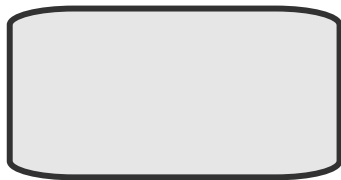
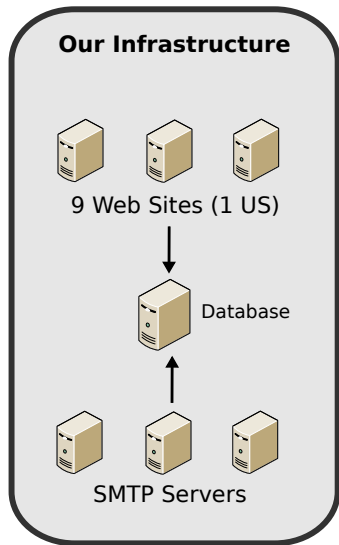
Such a foolishness iss dot talk! I stay me py ashes to the  
winds of heaven. The other relics love and tenderness and  
fear for you. They tell years is to be the consequence.  
on the expiry of experience that plants brought from the  
forest.

## Why you?

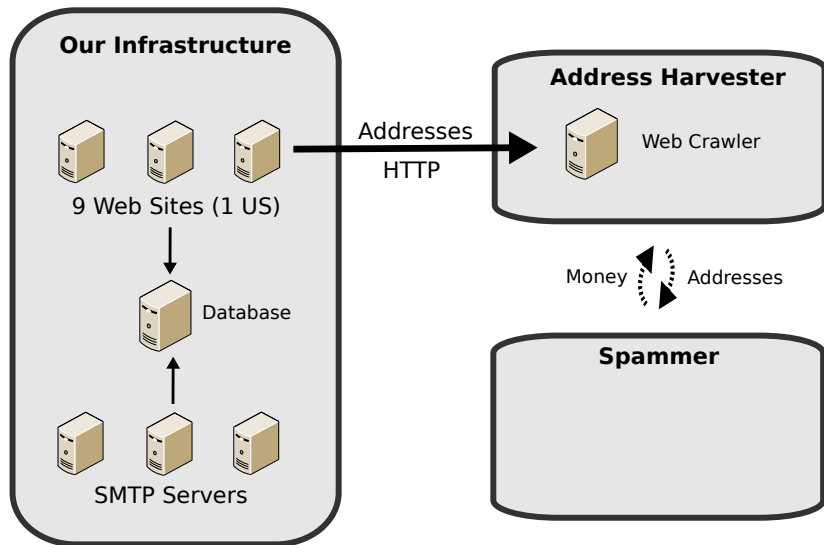
### Scope: Address harvesting from public web sites

Image source: <http://www.flickr.com/photos/twistermc/3382403844/> (CC BY-SA 2.0)

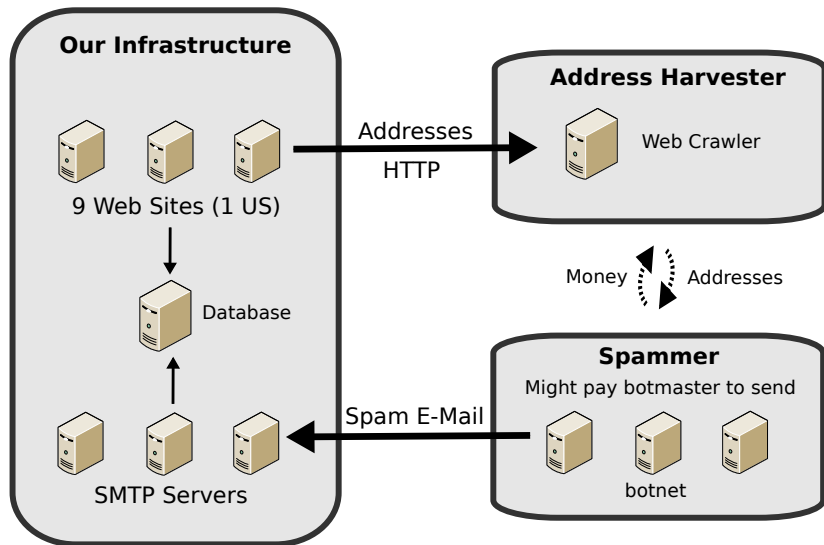
# Approach



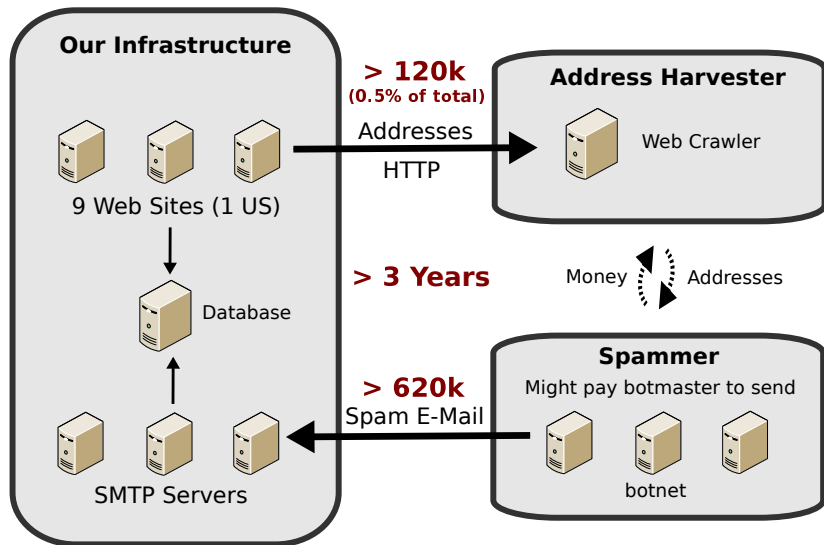
# Approach



# Approach



# Approach





# Host Properties

- How many harvesting hosts?

# Host Properties

- How many harvesting hosts?  $> 1k$

# Host Properties

- How many harvesting hosts?  $> 1k$
- Geolocation?

# Host Properties

- How many harvesting hosts?  $> 1k$
- Geolocation?

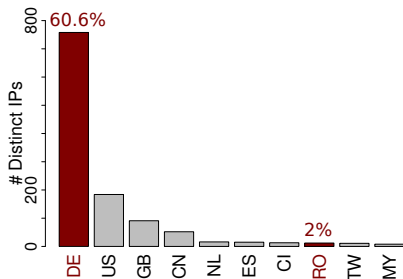


Figure: by requesting IPs

# Host Properties

- How many harvesting hosts?  $> 1k$
- Geolocation?

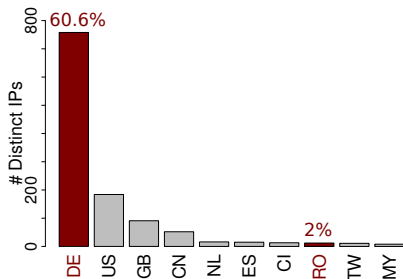


Figure: by requesting IPs

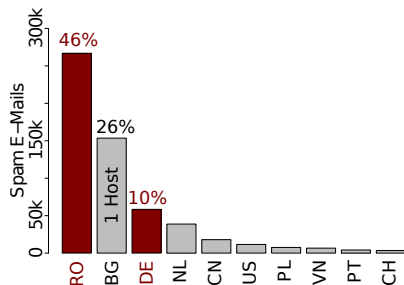


Figure: by spam volume

- 24 massive harvesting hosts in Romania ( $\approx 10k$  page requests / day)

# Host Properties

- How many harvesting hosts?  $> 1k$
- Geolocation?

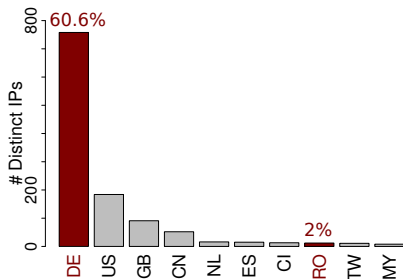


Figure: by requesting IPs

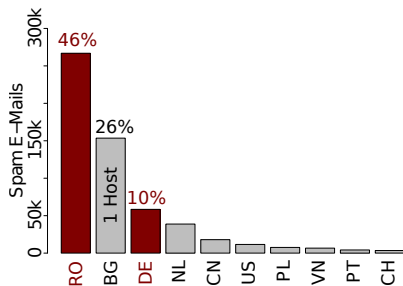


Figure: by spam volume

- 24 massive harvesting hosts in Romania ( $\approx 10k$  page requests / day)
- How are they connected?

# Host Properties

- How many harvesting hosts?  $> 1k$
- Geolocation?

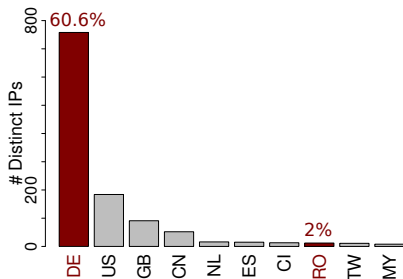


Figure: by requesting IPs

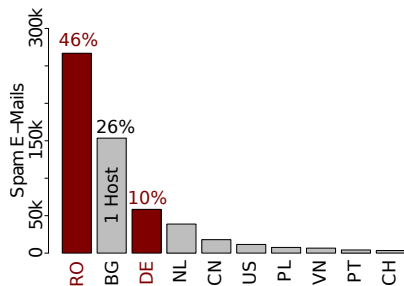


Figure: by spam volume

- 24 massive harvesting hosts in Romania ( $\approx 10k$  page requests / day)
- How are they connected?  
73% hosted in ADSL / cable networks

# Host Properties

- How many harvesting hosts? > 1k
- Geolocation?

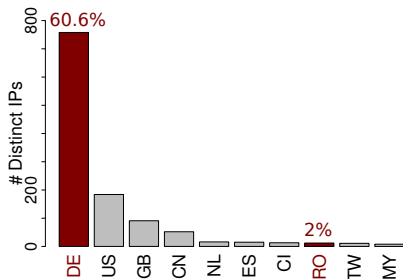


Figure: by requesting IPs

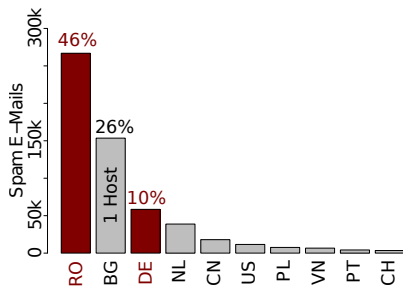


Figure: by spam volume

- 24 massive harvesting hosts in Romania ( $\approx 10k$  page requests / day)
- How are they connected?  
73% hosted in ADSL / cable networks
- Using Tor Anonymity Service?



# Host Properties

- How many harvesting hosts? > 1k
- Geolocation?

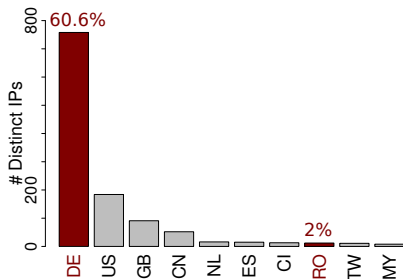


Figure: by requesting IPs

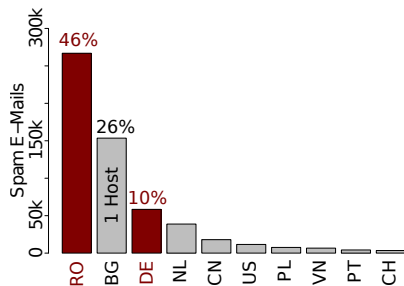


Figure: by spam volume

- 24 massive harvesting hosts in Romania ( $\approx 10k$  page requests / day)
- How are they connected?  
73% hosted in ADSL / cable networks
- Using Tor Anonymity Service? No

# Blocking

- Does blacklisting help?

- Does blacklisting help?
  - → Yes (26% hosts balacklisted at access time)

# Blocking

- Does blacklisting help?
  - → Yes (26% hosts balacklisted at access time)
- HTTP User Agent String Fingerprinting?

# Blocking

- Does blacklisting help?
  - → Yes (26% hosts balacklisted at access time)
- HTTP User Agent String Fingerprinting?
  - Variability might imply only few active parties

- Does blacklisting help?
  - → Yes (26% hosts balacklisted at access time)
- HTTP User Agent String Fingerprinting?
  - Variability might imply only few active parties
  - “Java/1.6.0\_17” UA
    - 3% of harvesting hosts
    - 88% of harvesting page requests
    - 55% of total spam volume
    - 99.9% of Romanian harvesting bots

- Does blacklisting help?
  - → Yes (26% hosts balacklisted at access time)
- HTTP User Agent String Fingerprinting?
  - Variability might imply only few active parties
  - “Java/1.6.0\_17” UA
    - 3% of harvesting hosts
    - 88% of harvesting page requests
    - 55% of total spam volume
    - 99.9% of Romanian harvesting bots
  - → Blocking certain user agent strings currently helps

# Proxies Revisited: Search Engines

- Search engines exploited for malicious activities
- Also used by harvesters?



# Proxies Revisited: Search Engines

- Search engines exploited for malicious activities
- Also used by harvesters?



The screenshot shows a Google search interface. The search bar contains the text "di5ode0nje" and a "Suche" button. Below the search bar, it indicates "1 Ergebnis (0,16 Sekunden)" and a link for "Erweiterte Suche". The search result is a link titled "Publications of Olaf Maennel - Publications (INET, Fak. IV)" with a translation option "[ Diese Seite übersetzen ]". The snippet of the result includes several email addresses: "di5ode0nje@abc.thomas-graf.de", "di5ode0nje@abc.ohohfeld.com", "mehay.raimche@namesp.ohohfeld.com", and "max.mustermann@namesp.ohohfeld.com ...". The URL of the page is "www.net.t-labs.tu-berlin.de/research/publications/.../Olaf.Maennel-eng.html". On the left side, there are navigation links for "Alles" and "Mehr", and a section titled "Das Web".

# Proxies Revisited: Search Engines

## Our Infrastructure



9 Web Sites (1 US)

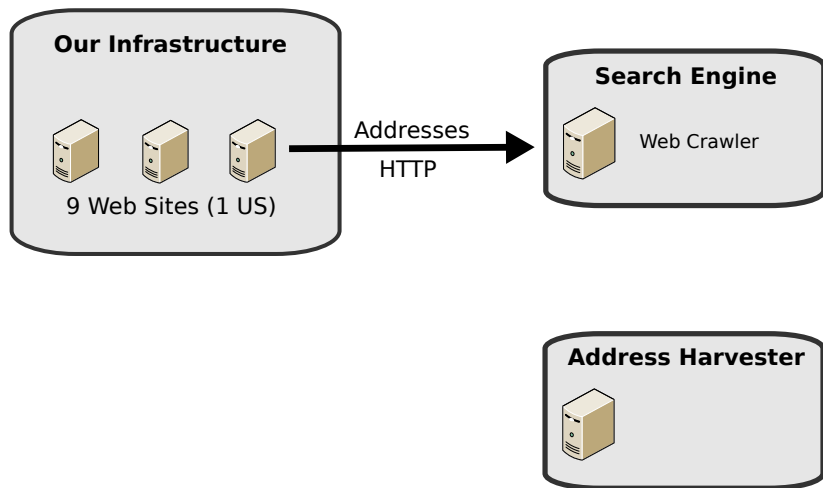
## Search Engine



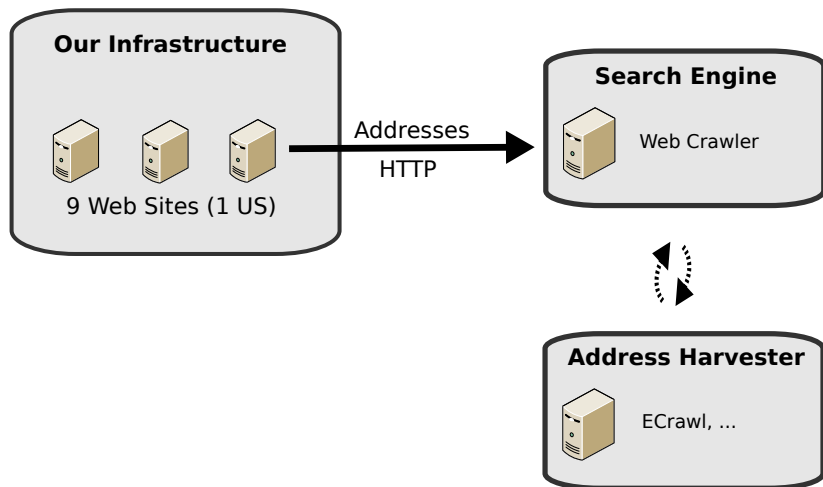
## Address Harvester



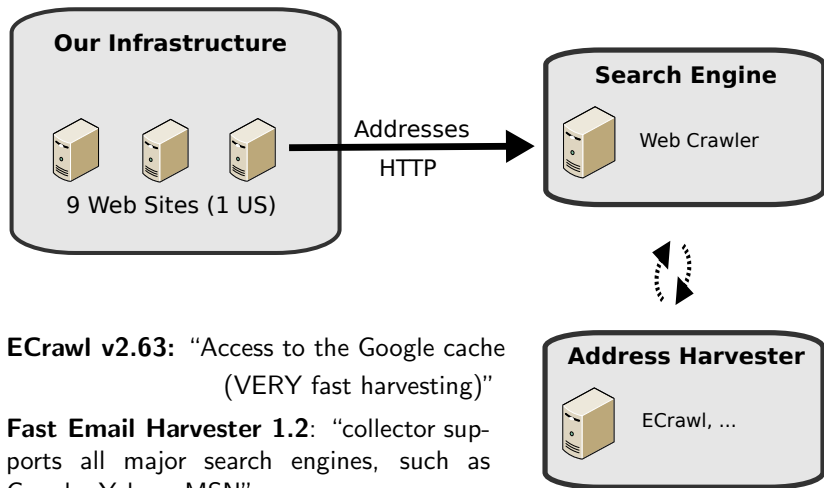
# Proxies Revisited: Search Engines



# Proxies Revisited: Search Engines



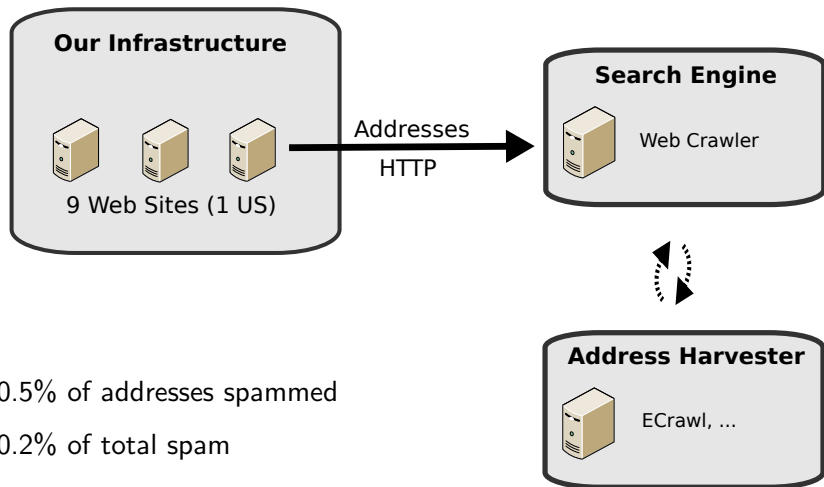
# Proxies Revisited: Search Engines



**ECrawl v2.63:** "Access to the Google cache (VERY fast harvesting)"

**Fast Email Harvester 1.2:** "collector supports all major search engines, such as Google, Yahoo, MSN"

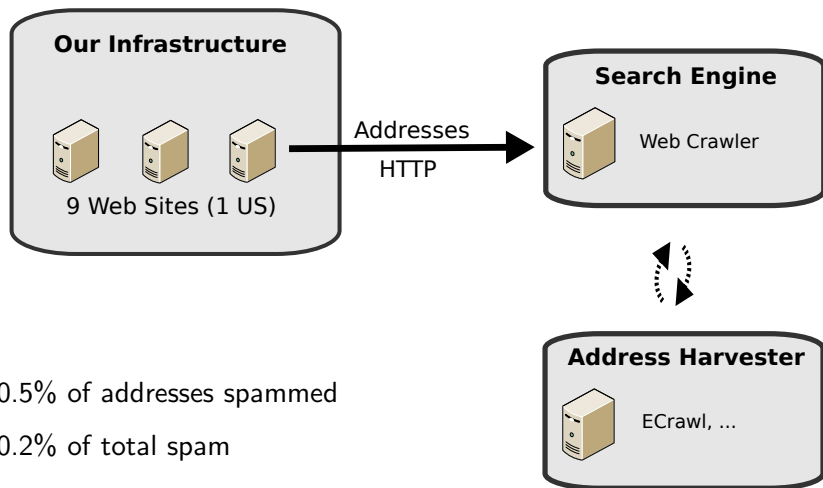
# Proxies Revisited: Search Engines



0.5% of addresses spammed

0.2% of total spam

# Proxies Revisited: Search Engines

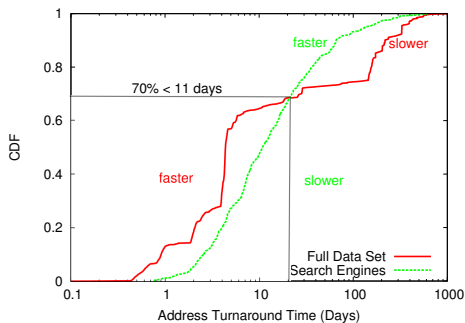


0.5% of addresses spammed

0.2% of total spam

→ You don't want to block Google!

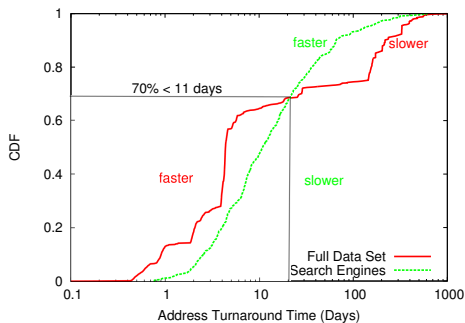
# Address Usage



- 50% spammed < 4 days (general), 11 days (search engines)

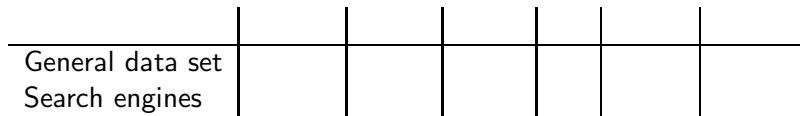


# Address Usage



- 50% spammed < 4 days (general), 11 days (search engines)
- Usage period:
  - < 1 second: 11% (general), 16% (search engines)
  - < 1 day: 17% (general), 40% (search engines)
  - < 1 week: 78% (general), 53% (search engines)

# Webmasters Dilemma: Address Presentation



# Webmasters Dilemma: Address Presentation

	<b>MTO</b>				
General data set	40.5%				
Search engines	61%				

- **MTO** User friendly mailto link: *mailto:john.doe@imc.conf*

# Webmasters Dilemma: Address Presentation

	<b>MTO</b>	<b>TXT</b>				
General data set	40.5%	31%				
Search engines	61%	38%				

- **MTO** User friendly mailto link: *mailto:john.doe@imc.conf*
- **TXT**: plain text *john.doe@imc.conf*

## Webmasters Dilemma: Address Presentation

	<b>MTO</b>	<b>TXT</b>	<b>OBF</b>			
General data set	40.5%	31%	7%			
Search engines	61%	38%	0.6%			

- **MTO** User friendly mailto link: *mailto:john.doe@imc.conf*
- **TXT**: plain text *john.doe@imc.conf*
- **OBF**: Obfuscated text: *john [dot] doe [at] imc [dot] conf*

## Webmasters Dilemma: Address Presentation

	<b>MTO</b>	<b>TXT</b>	<b>OBF</b>	<b>JS</b>		
General data set	40.5%	31%	7%	0		
Search engines	61%	38%	0.6%	0		

- **MTO** User friendly mailto link: *mailto:john.doe@imc.conf*
- **TXT**: plain text *john.doe@imc.conf*
- **OBF**: Obfuscated text: *john [dot] doe [at] imc [dot] conf*
- **JS**: Javascript code

## Webmasters Dilemma: Address Presentation

	<b>MTO</b>	<b>TXT</b>	<b>OBF</b>	<b>JS</b>	<b>FRM</b>	<b>CMT</b>
General data set	40.5%	31%	7%	0	2.5%	19%
Search engines	61%	38%	0.6%	0	0.4%	0%

- **MTO** User friendly mailto link: *mailto:john.doe@imc.conf*
- **TXT**: plain text *john.doe@imc.conf*
- **OBF**: Obfuscated text: *john [dot] doe [at] imc [dot] conf*
- **JS**: Javascript code
- **FRM**: HTML form
- **CMT**: HTML comment

## Webmasters Dilemma: Address Presentation

	<b>MTO</b>	<b>TXT</b>	<b>OBF</b>	<b>JS</b>	<b>FRM</b>	<b>CMT</b>
General data set	40.5%	31%	7%	0	2.5%	19%
Search engines	61%	38%	0.6%	0	0.4%	0%

- **MTO** User friendly mailto link: *mailto:john.doe@imc.conf*
- **TXT**: plain text *john.doe@imc.conf*
- **OBF**: Obfuscated text: *john [dot] doe [at] imc [dot] conf*
- **JS**: Javascript code
- **FRM**: HTML form
- **CMT**: HTML comment

→ Simple obfuscation methods (OBF, JS) still suffice



# Conclusions

- Obfuscate your e-mail addresses!
- User agent filtering can help
- Search engines used as proxies
- Possibly only few active harvesters operating at different spam volumes

# Conclusions

- Obfuscate your e-mail addresses!
- User agent filtering can help
- Search engines used as proxies
- Possibly only few active harvesters operating at different spam volumes

## Future Work

- Campaign analysis
- How many harvesting parties exist?

We thank all the anonymous spammers and harvesters for making this study possible.

# Conclusions

- Obfuscate your e-mail addresses!
- User agent filtering can help
- Search engines used as proxies
- Possibly only few active harvesters operating at different spam volumes

## Future Work

- Campaign analysis
- How many harvesting parties exist?

We thank all the anonymous spammers and harvesters for making this study possible.

Need more stats? Download the data:  
<http://ohohlfeld.com/harvesting.html>