Empirical Analysis of Factors Affecting Malware URL Detection

Marie Vasek & Tyler Moore
SMU, Dallas, TX
mvasek@smu.edu

eCrime Researchers Summit

September 17, 2013

Table of Contents

So Much Malware

So Much Malware

So Many AntiVirus Products

Research Questions

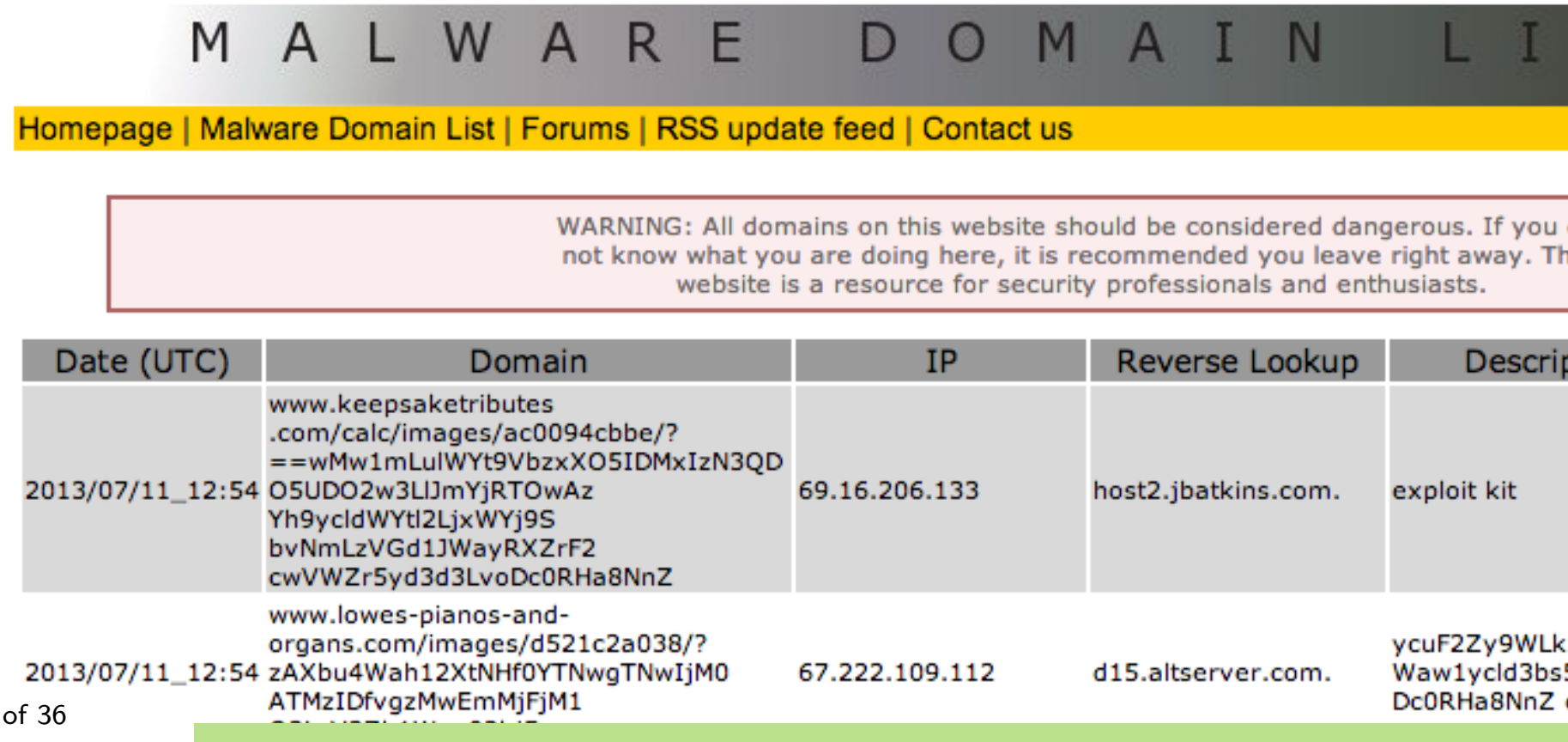
Our hypothesis: At least part of the difference in performance across blacklists can be explained by characteristics of the malware URL or the blacklisting service.

1. Impact of type of malware URL on blacklisting?
   - Payload type (exploit kit, Fake AV, etc.)
   - URL features (IP address, path)
2. Impact of type of blacklisting service?
   - Blocking or advisory
   - Cost

Table of Contents

Malware Domain List

- Publicly Available
- Low Volume
  - 722 URLs in our collection
  - *BUT* it's free, open, & no data-sharing issues



M A L W A R E D O M A I N L I

Homepage | Malware Domain List | Forums | RSS update feed | Contact us

WARNING: All domains on this website should be considered dangerous. If you
not know what you are doing here, it is recommended you leave right away. Th
website is a resource for security professionals and enthusiasts.

Date (UTC)	Domain	IP	Reverse Lookup	Descrip
2013/07/11_12:54	www.keepsaketributes .com/calc/images/ac0094cbbe/? ==wMw1mLulWYt9VbzxXO5IDMxIzN3QD O5UDO2w3LlJmYjRTOwAz Yh9ycldWYtl2LjxWYj9S bvNmLzVGd1JWayRXZrF2 cwVWZr5yd3d3LvoDc0RHa8NnZ	69.16.206.133	host2.jbatkins.com.	exploit kit
2013/07/11_12:54	www.lowes-pianos-and- organs.com/images/d521c2a038/? zAXbu4Wah12XtNHf0YTNwgTNwIjM0 ATMzIDfvgzMwEmMjFjM1	67.222.109.112	d15.altserver.com.	ycuF2Zy9WLk Waw1ycld3bs5 Dc0RHa8NnZ

VirusTotal

- Publicly Available
- API accessible
- Google-Owned
- Checks each URL against 38 blacklists

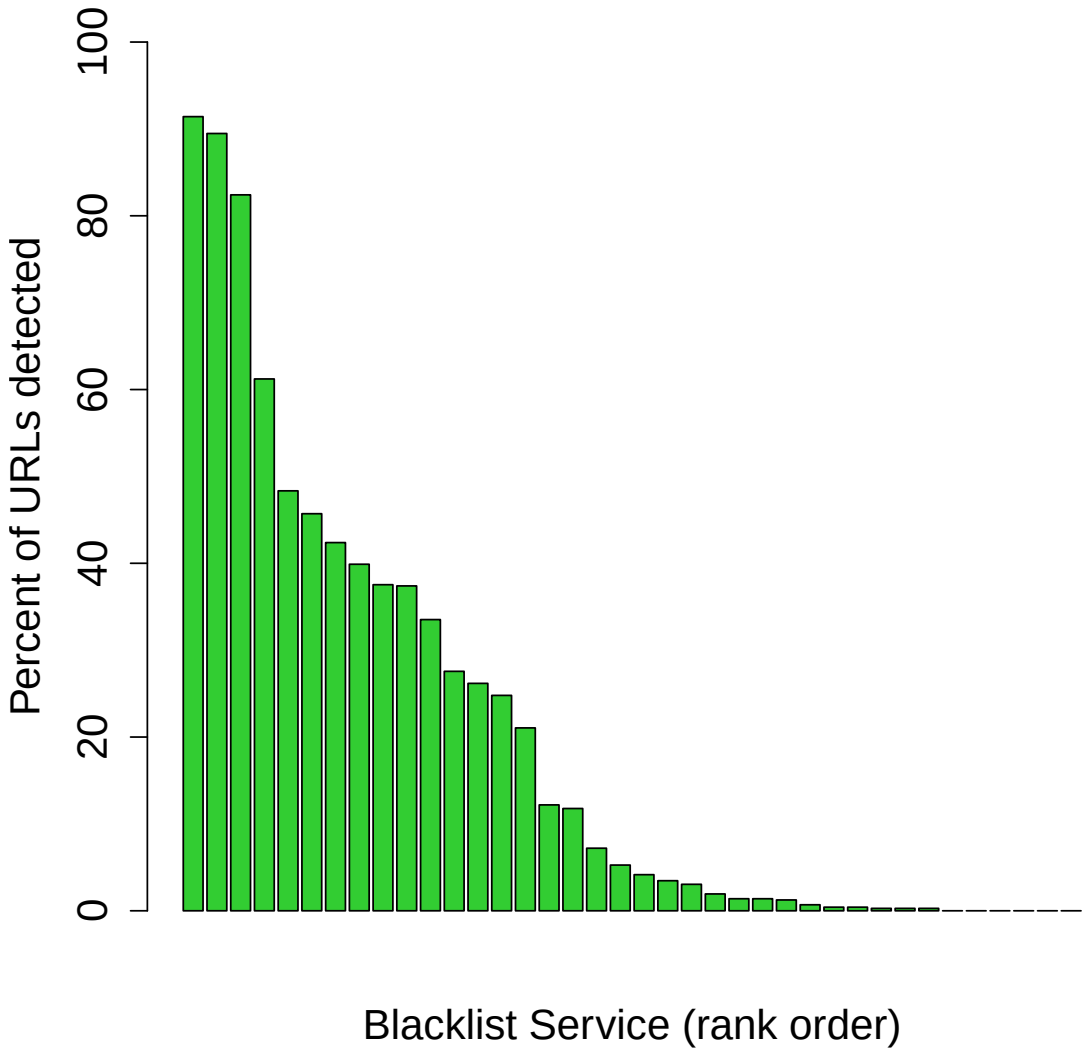
Table of Contents

What We Do

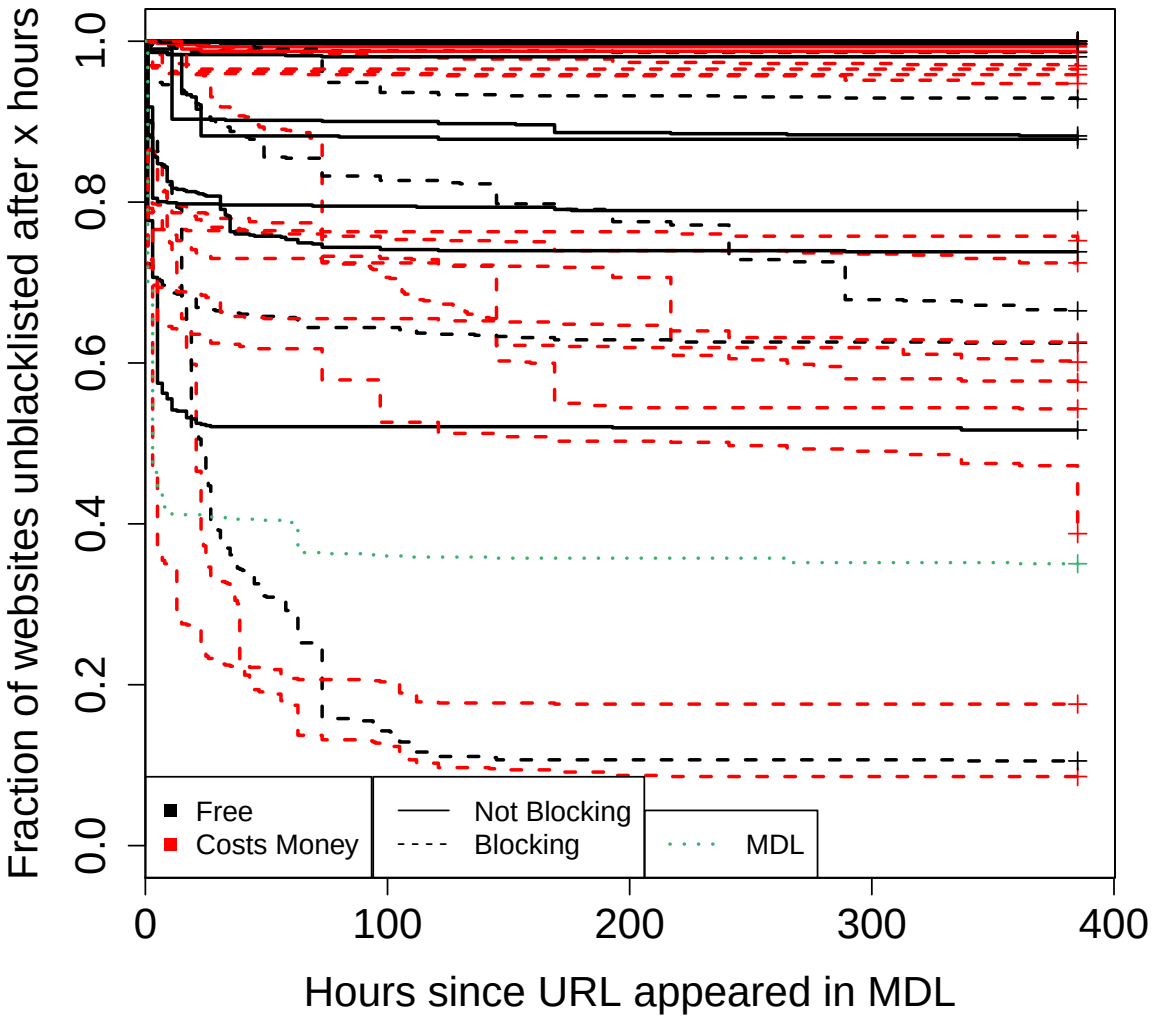
1. Check MDL for new URLs.
2. Scan new URLs through VT.
3. Scan new URL every other hour for 48 hours after.
4. Afterwards, scan every day for 14 days.

Table of Contents

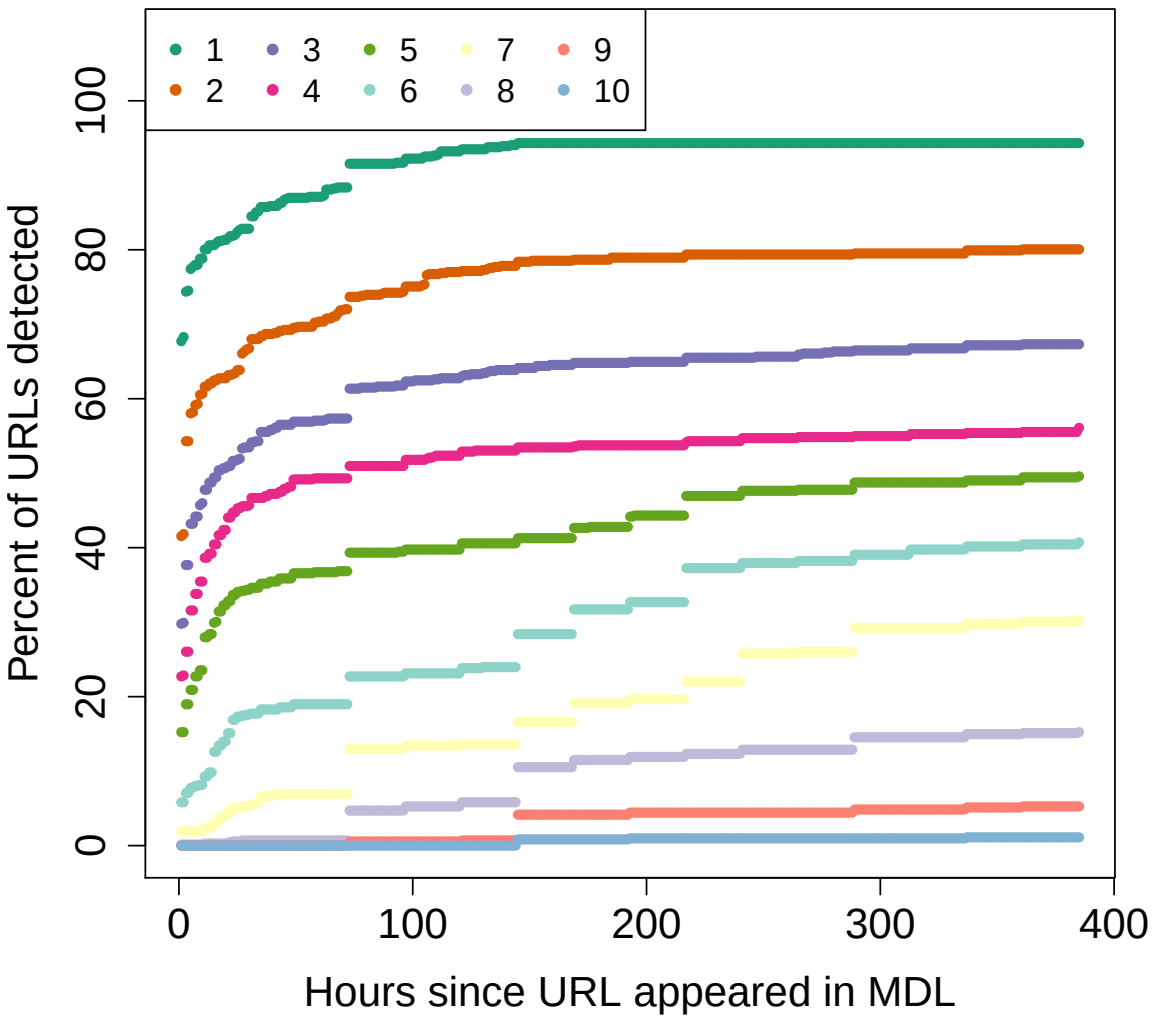
How Blacklisting Detection Rates Vary



How Blacklisting Detection Rates Vary



Percentage of URLs Detected by at least *n* Services



Blacklisting over Time

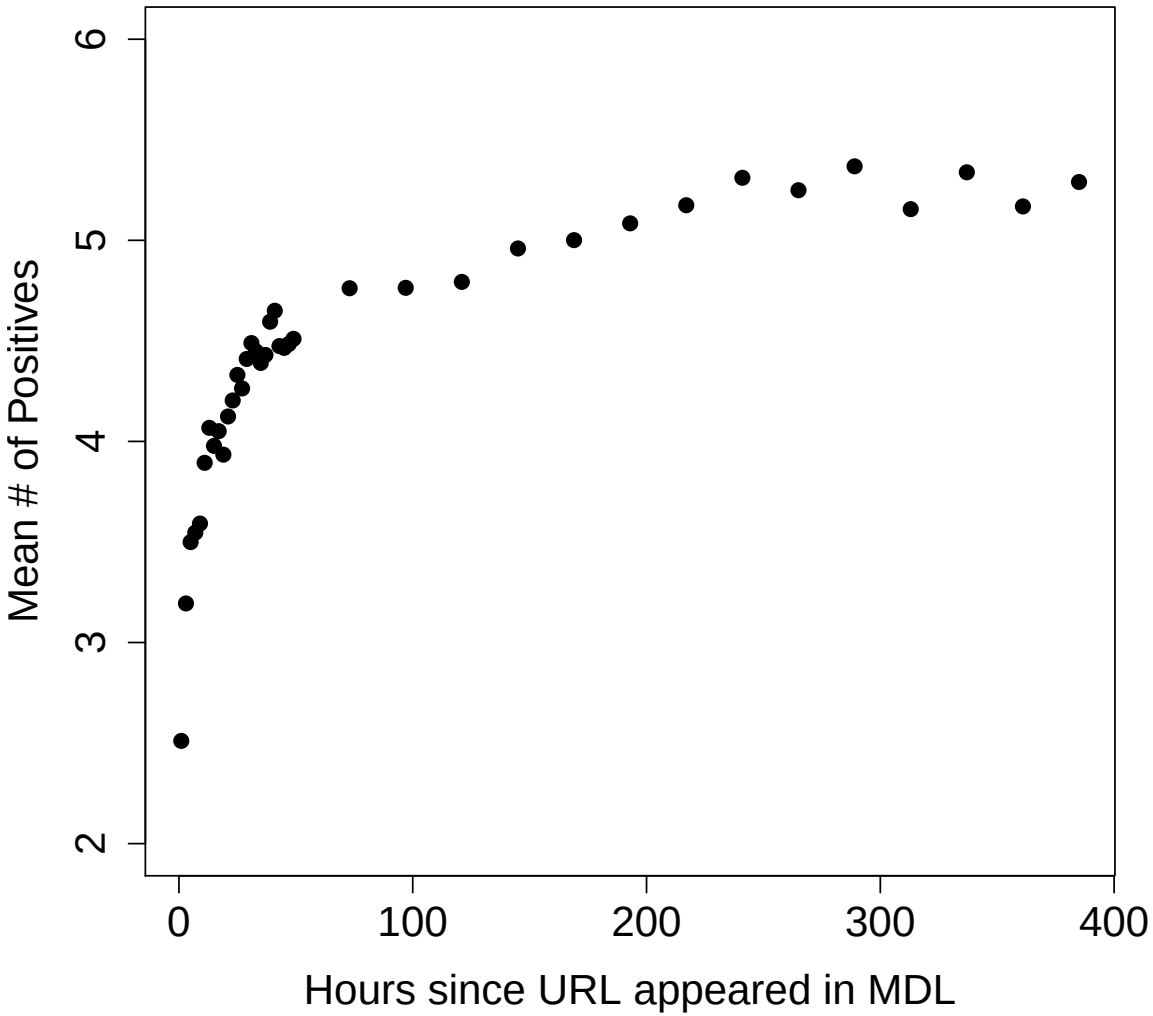


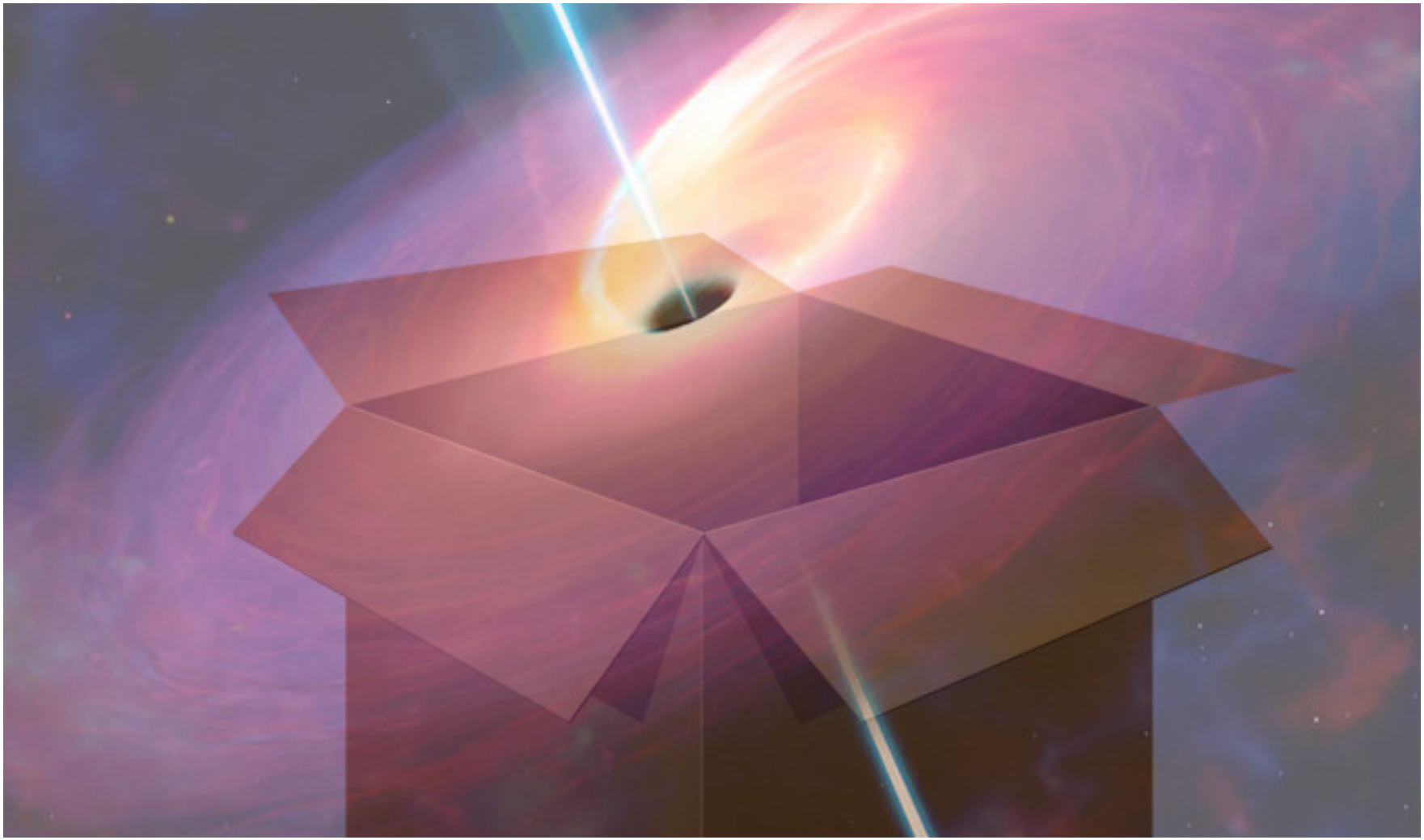
Table of Contents

Research Questions

Our hypothesis: At least part of the difference in performance across blacklists can be explained by characteristics of the malware URL or the blacklisting service.

1. Impact of type of malware URL on blacklisting?
   - Payload type (exploit kit, Fake AV, etc.)
   - URL features (IP address, path)
2. Impact of type of blacklisting service?
   - Blocking or advisory
   - Cost

Malware Type



IP/Domain

127.0.0.1/bad.html

icanhazcats.com/milk.php

Path?

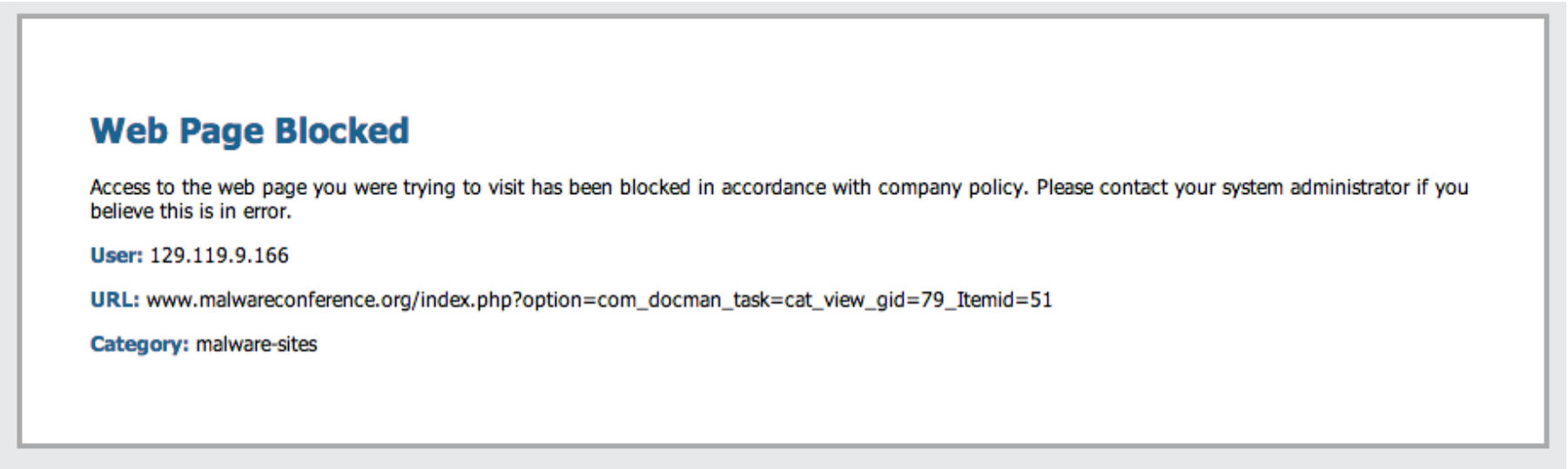
fluffybunnies.org/

goldphish.net/bubbles.php

Type of Malware

Malware Type	#	%	IP/Domain	#	%	Path?	#	%
Executable	175	24%	IP Address	124	17%	Has Path	675	93%
Fake AV	65	9%	Domain	598	83%	No Path	47	7%
Styx	51	7%						
Blackhole Lnd.	149	21%						
Other	282	39%						

Blocks?



Web Page Blocked

Access to the web page you were trying to visit has been blocked in accordance with company policy. Please contact your system administrator if you believe this is in error.

User: 129.119.9.166

URL: www.malwareconference.org/index.php?option=com_docman_task=cat_view_gid=79_Itemid=51

Category: malware-sites

Costs Money

Blacklist Types

Blocks?	Costs?	Description
✓	✓	Paid AV
✓	✗	Free AV & Web Browser Blacklists
✗	✓	Malware Site Checkers
✗	✗	Malware URL lists

Blacklist Types

Blocks?	#	%	Costs?	#	%
Blocks Users	22	58%	Costs Money	17	45%
Doesn't Block	16	42%	Free	21	55%

Table of Contents

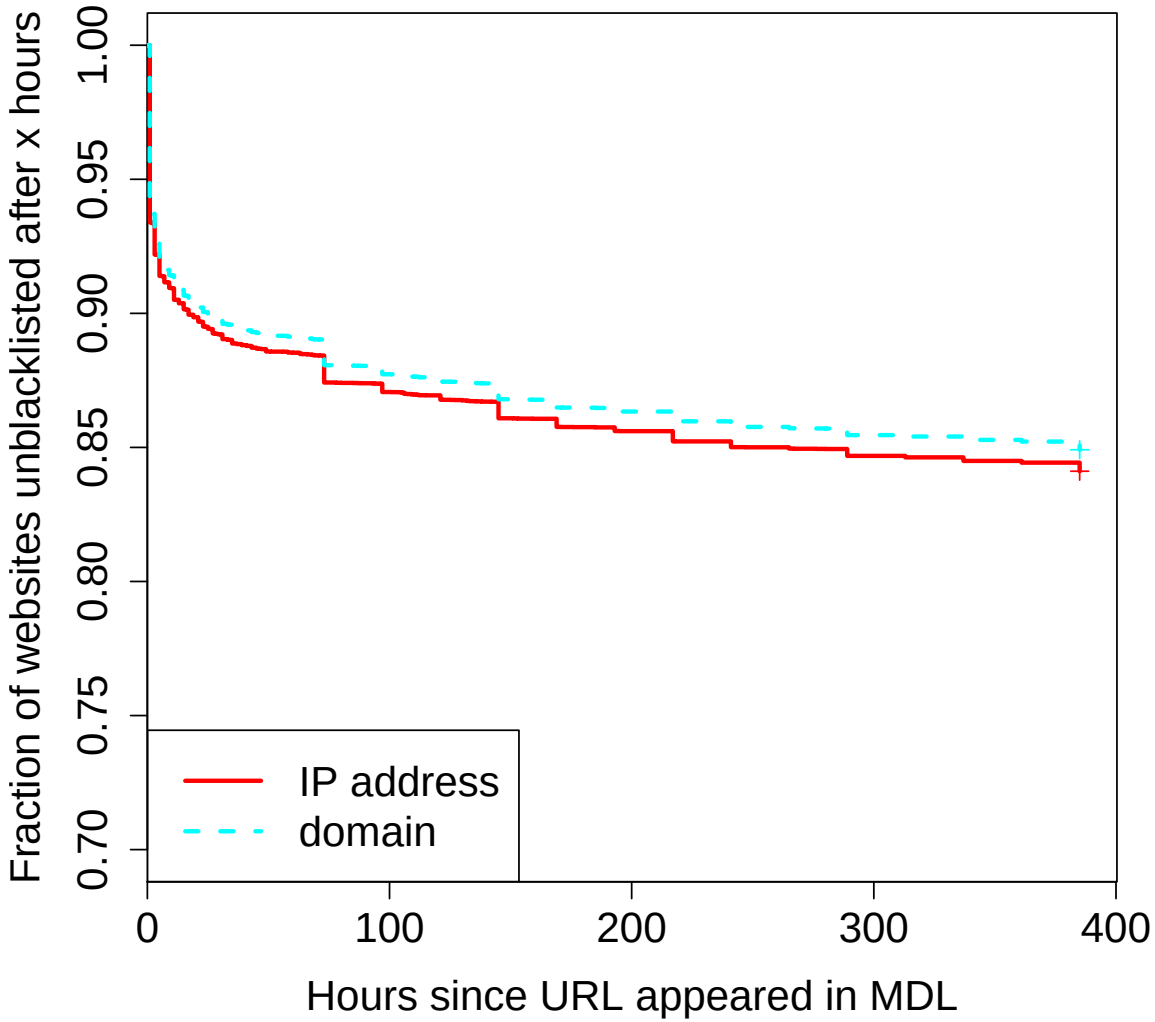
Logistic Regression: URL ever Blacklisted by Service?

	Coefficient	Odds Ratio	p value
Intercept	-2.304	**0.100**	0.000
URL features			
IP address	-0.004	0.996	0.945
Has a Path?	-0.196	**0.822**	0.031
Executable	1.017	**2.765**	0.000
Fake AV	-0.838	**0.433**	0.000
Styx	0.746	**2.109**	0.000
Blackhole Landing Page	0.196	**1.217**	0.000
Malware Blacklist Features			
Blocks Users?	0.611	**1.843**	0.000
Costs Money	0.298	**1.347**	0.000
$\chi^2 = 1144.501$, p value $= 0.000$			

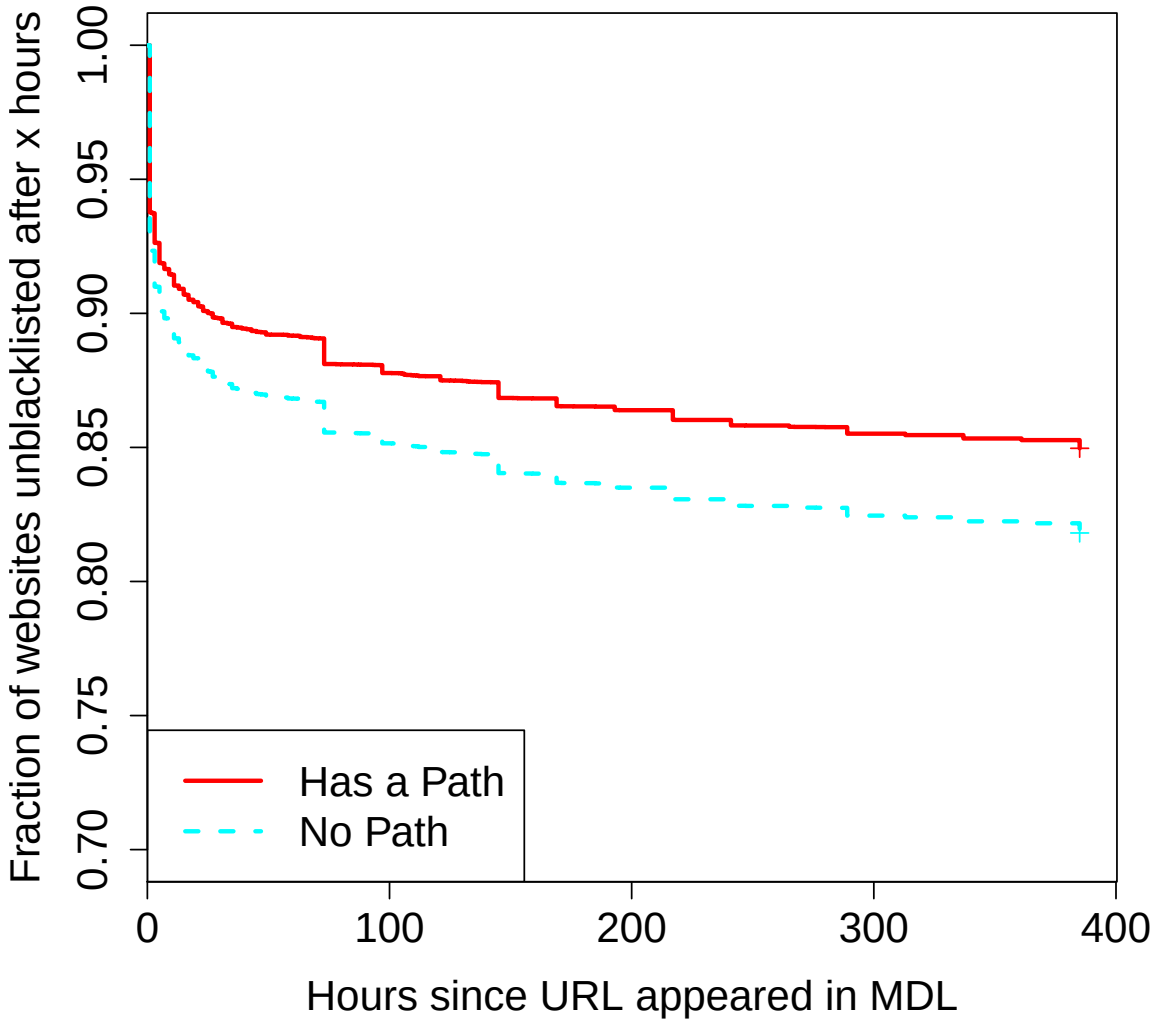
Survival Regression: Blacklist Timing

	Coefficient	Odds Ratio	p value
URL features			
IP address	0.056	1.058	0.210
Has a Path?	-0.207	**0.811**	0.012
Executable	0.896	**2.449**	0.000
Fake AV	-0.814	**0.443**	0.000
Styx	0.750	**2.118**	0.000
Blackhole Landing Page	0.179	**1.196**	0.000
Malware Blacklist Features			
Blocks Users?	0.538	**1.713**	0.000
Costs Money	0.300	**1.351**	0.000
$R^2 = 0.055$			

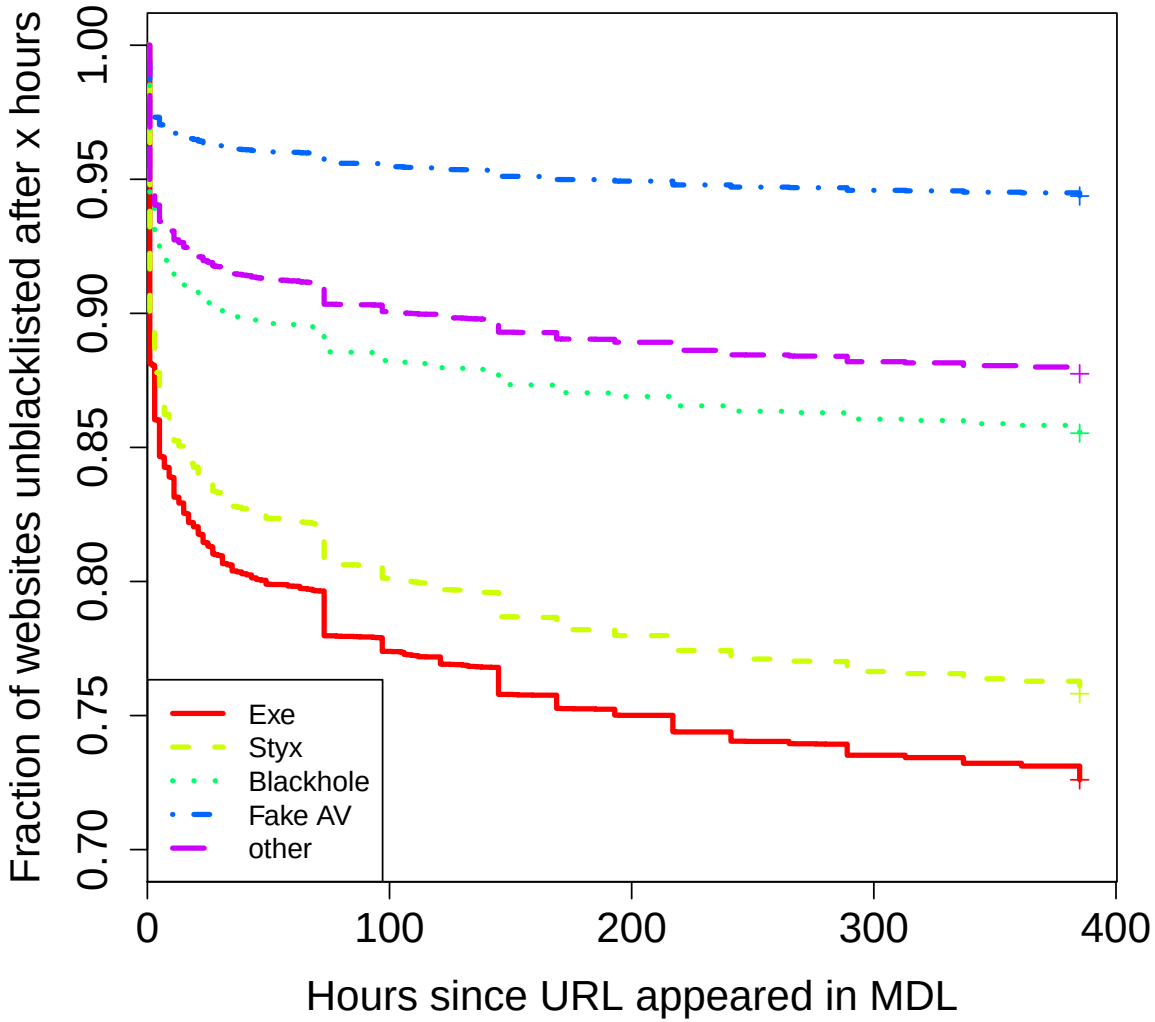
Survival Regression



Survival Regression



Survival Regression



Survival Regression

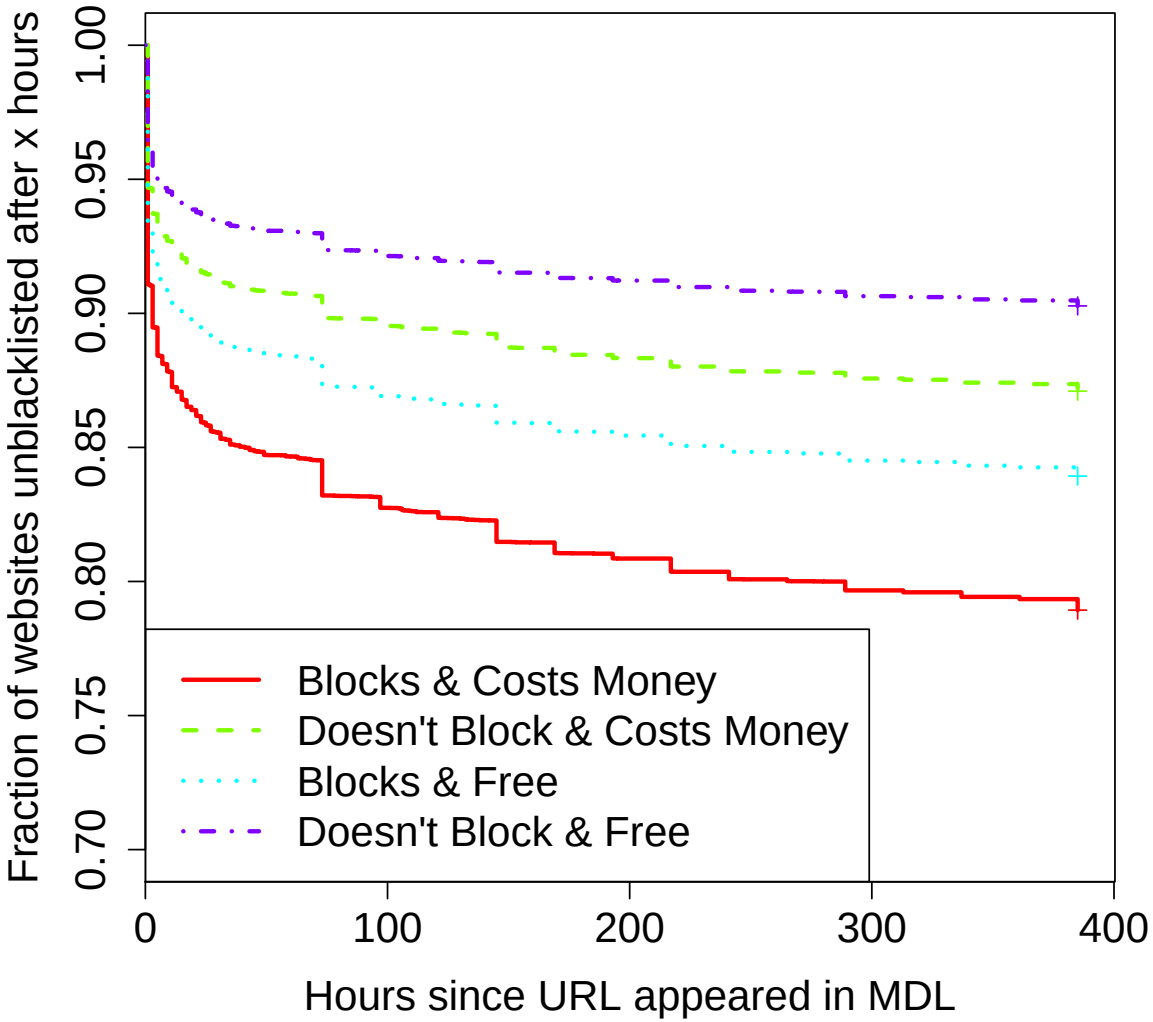


Table of Contents

Summary

- Known Exploit kits more likely to be blacklisted
  - ... but Fake AV is not ..?
- Executables more likely to be blacklisted
- Blocking users $\longrightarrow$ more likely to blacklist
- Costing money $\longrightarrow$ more likely to blacklist

Limitations

- MDL $\longrightarrow$ low volume
  - other exploit kits
  - AS / DNS features
  - Malicious vs. Hacked?
- MDL $\longrightarrow$ used as input to blacklists
  - Blacklisted b/c found on VT or found organically?
  - Are things on MDL still bad on day n?
- VT $\longrightarrow$ not representative sample of malware blacklists
  - Blacklists not on VT?
  - Blacklists on VT that didn't blacklist anything?

Questions?