The Long Story of Short URLs

Federico Maggi
Politecnico di Milano

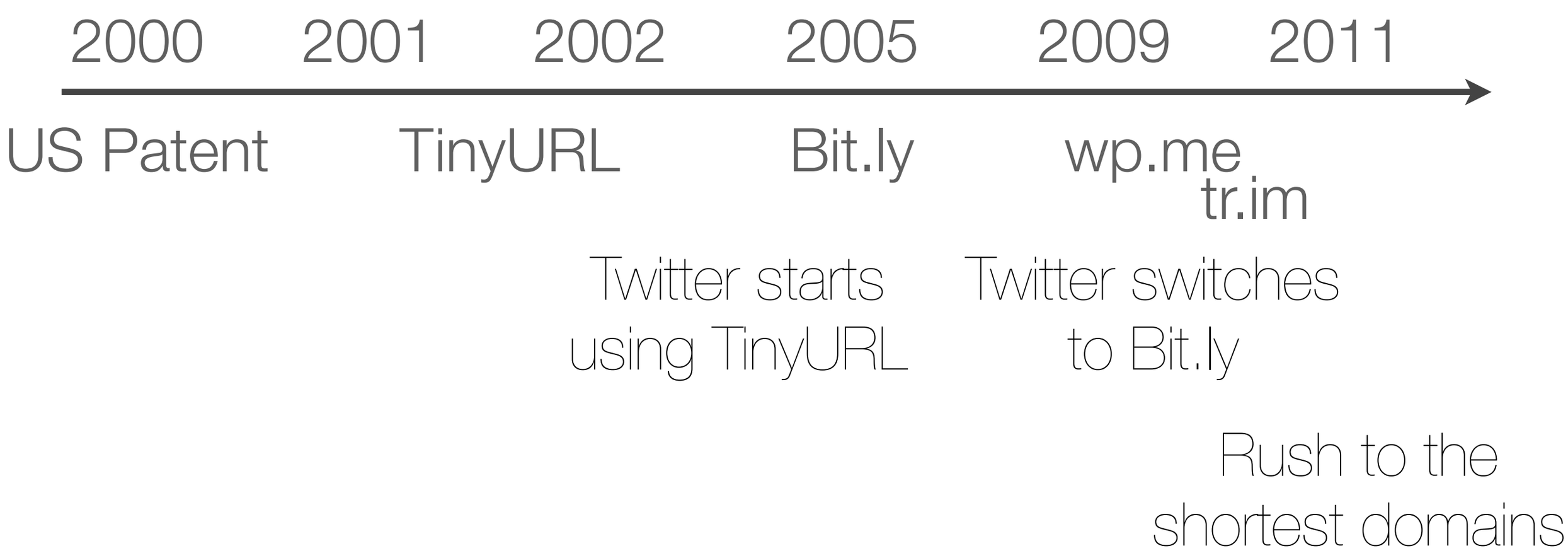
http://www.syssec-project.eu

syssec European System Security Researchers

- The **research** leading to the **results presented in this talk** has received **funding** from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 257007.

- Builds on the FORWARD initiative, SysSec **aims** at:
  - creating a virtual center of **excellence**, to **consolidate** the systems security research community in **Europe**,
  - promoting cybersecurity **education,**
  - engaging a **think-tank** in discovering the threats and vulnerabilities,
  - creating an active research **roadmap** in the area, and
  - developing a joint working plan to conduct **collaborative research**.

Brief history of short URLs

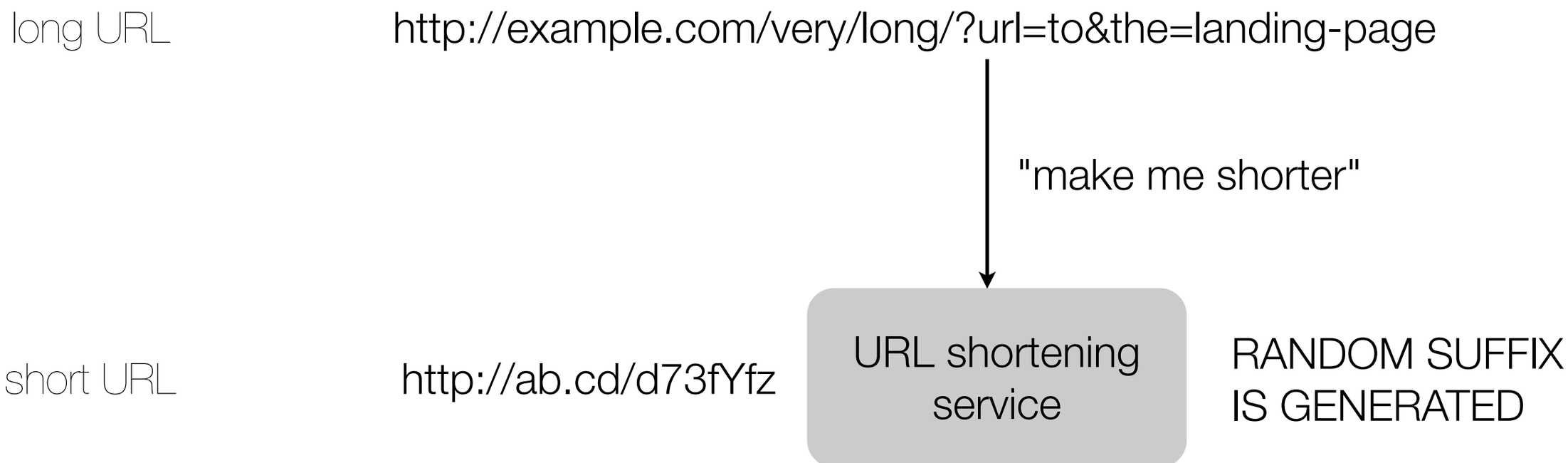


Today is it just bit.ly and t.co?

- We observed up to **622** shortening services
- **Companies** and famous **bloggers** have started using their own custom domains (e.g., pep.si, ti.me, flic.kr)

Short URLs have become a sort of "**trendy gadget**"

How short URLs work



How short URLs work cont'd

REDIRECTION
MECHANISM

http://ab.cd/d73fYfz → http://example.com/very/long/...

HTTP 302

HTML META Refresh

JavaScript

ActionScript

executes on the browser's side

http://ab.cd/d73fYfz ⟶ http://ab.cd/123fa1 ⟶ http://ab.cd/44a8F ⟶ http://ab.cd/as9fYc

Why short URLs could be misused

- **Users** have grown **accustomed** to see short URLs
- Users typically **trust** short URLs
- They look **harmless**

http://srv153.example.com/very/long/?url=to&the=landing-page&p=121&id=20&par=value&very=suspicious&long=url&that=would&probably=not&fit=into&your=IM&chat=window&or=may&be=broken&into=severla=lines

VS

http://i.am/so-tiny

From the bad guys' perspective

Perfect mean for **masquerading** suspicious URLs

- Trivially **evade** naïve checks
- **Trendy** effect (e.g., Twitter, Facebook)
- **Robust** to those clients that break long URLs into multiple lines
- **Dynamic** redirection mechanisms (e.g., JavaScript, timeout, "Click to continue") make the landing page unaccessible to **automated scanners**

State of the art and related work

- Spam, phishing and other malicious activity on **social networks** use short URLs
  - [Stringhini et al., ACSAC **2010**], [Grier et al., CCS 2010], [Gao et al., IMC **2010**]
- "Quality" of the **content** aliased via short URLs is either very high or very low
  - [Kandylas et al., WWW **2010**]
- **Crawling** existing short URLs and use APIs to expand and analyze them
  - [Antoniades et al., WWW **2011**]
- Common nodes of the **redirection chains** are distinctive of bad short URLs
  - [Lee, S. and Kim, J., NDSS **2012**]

These work consider existing short URLs **found** on websites

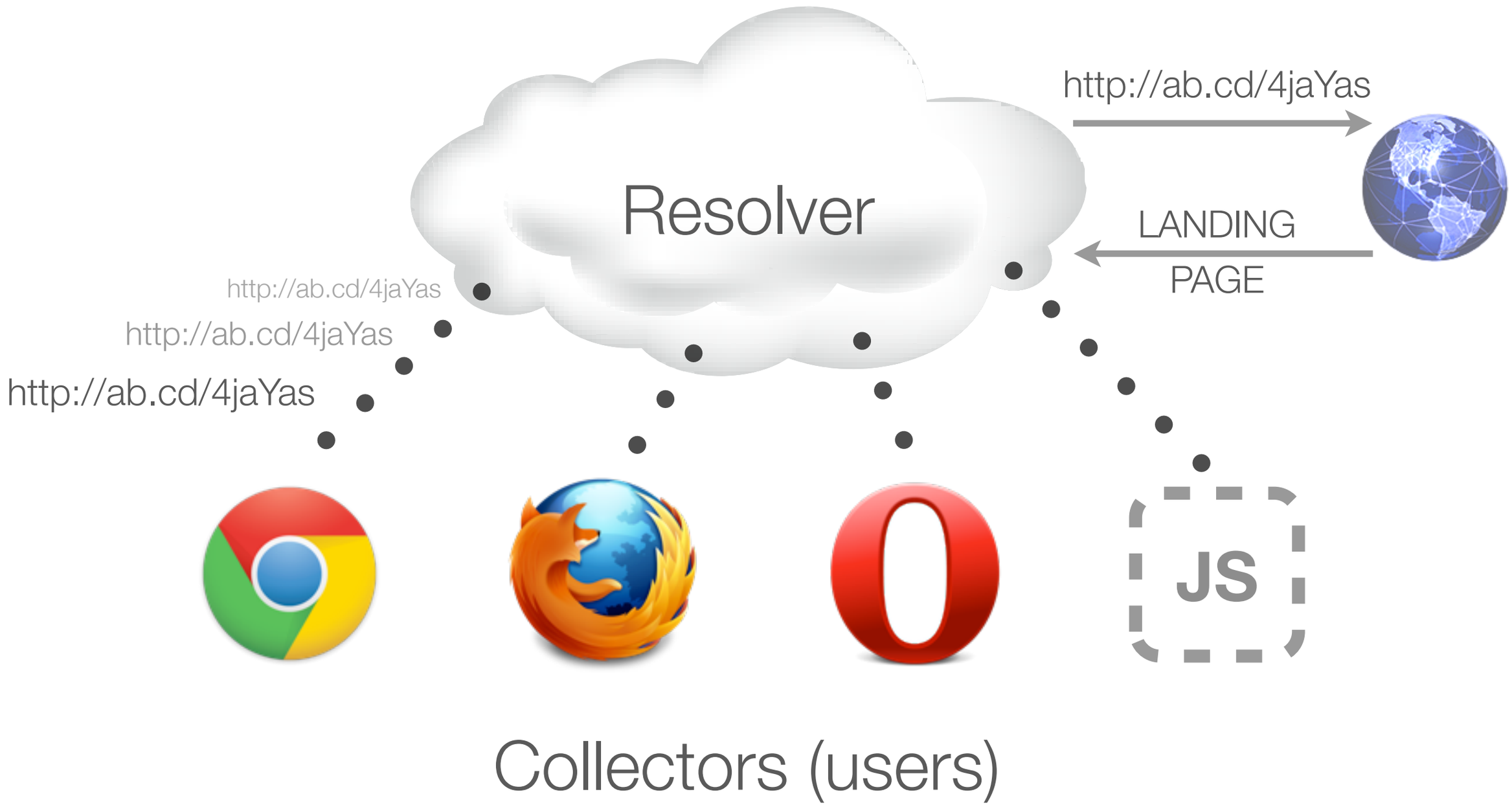
None of them take the **end users** into account

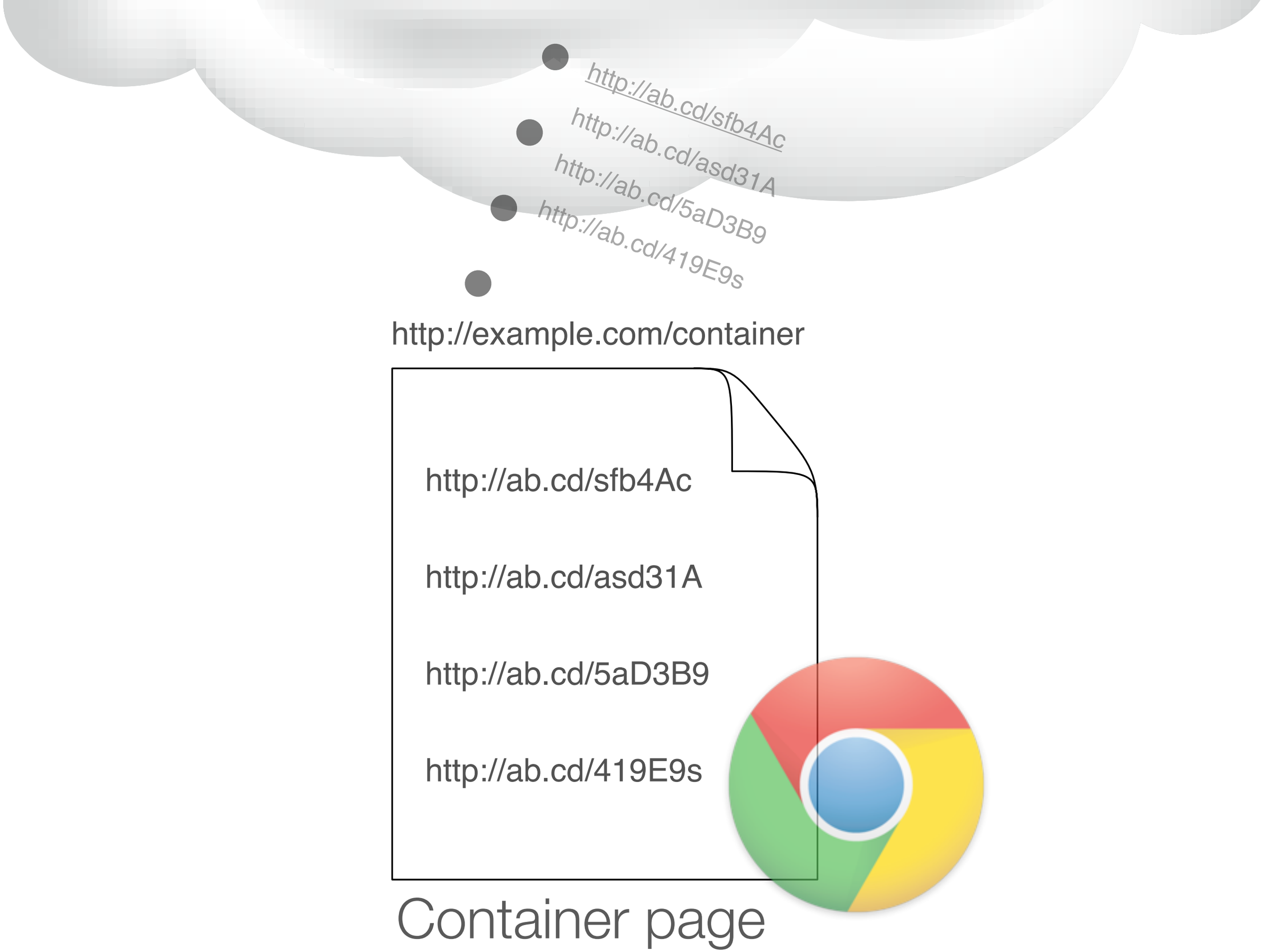
A different perspective what is the impact on users?

- What kind of short URLs **users** typically encounter?
- Do users stumble upon **malicious** short URLs that **often**?
- Do users **perceive** the maliciousness of a short URL?
- Do shortening services take enough **countermeasures** to protect the users?

User-centered measurement

Data collection infrastructure




http://ab.cd/sfb4Ac
http://ab.cd/asd31A
http://ab.cd/5aD3B9
http://ab.cd/419E9s
http://example.com/container
http://ab.cd/sfb4Ac
http://ab.cd/asd31A
http://ab.cd/5aD3B9
http://ab.cd/419E9s
Container page

How to avoid biased measurements?

- We do not ask a user to **become a collector**
- We provide a **useful service** that users may need
- Users **spontaneously** subscribe as collectors

What kind of **service** do we offer?

- http://bit.ly/N9NwC - points to Wikipedi
- http://tinyurl.com/2ks - points to Yahoo
- http://tiny.cc/txnfl - points to World Wi

your choice

World Wide Web Consortium (W3C)

Long URL: http://w3.org/
Destination: http://www.w3.org/
Size: 28490 bytes
Type: text/html
Details: see detailed analysis

+flag as malicious!

apis.com/o

com/api/x

Data that we collect

Raw data

Timestamp

Short URL

Client's IP

Referrer

Extracted data

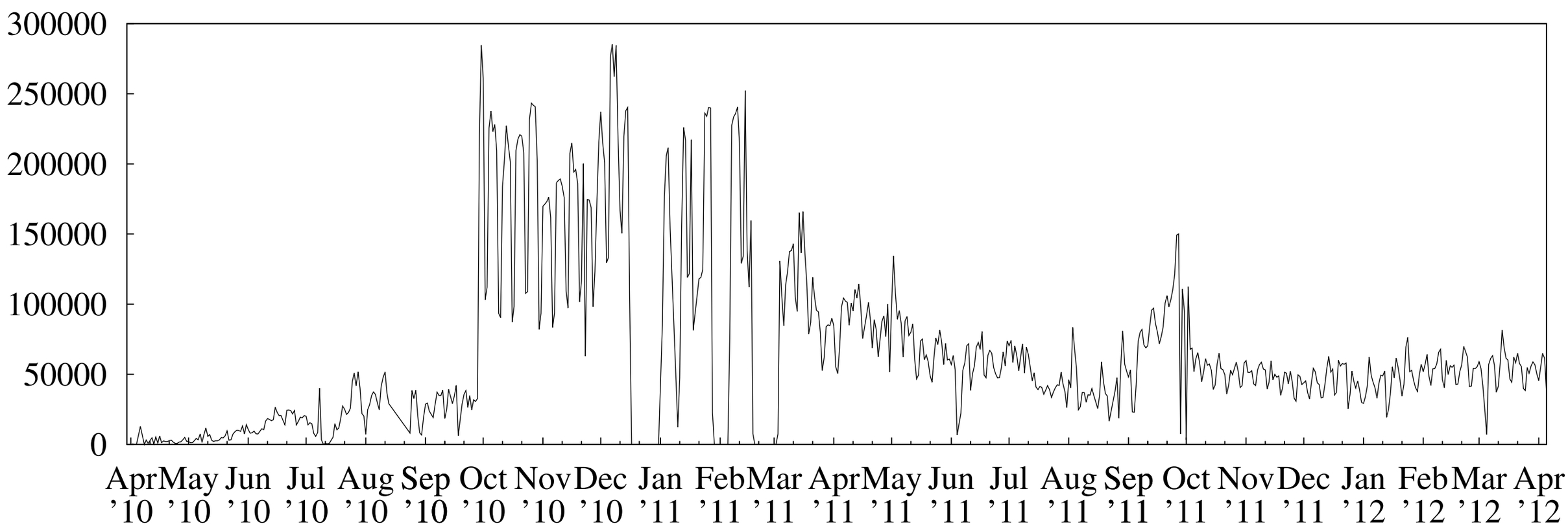
Next hop

Redirection chain

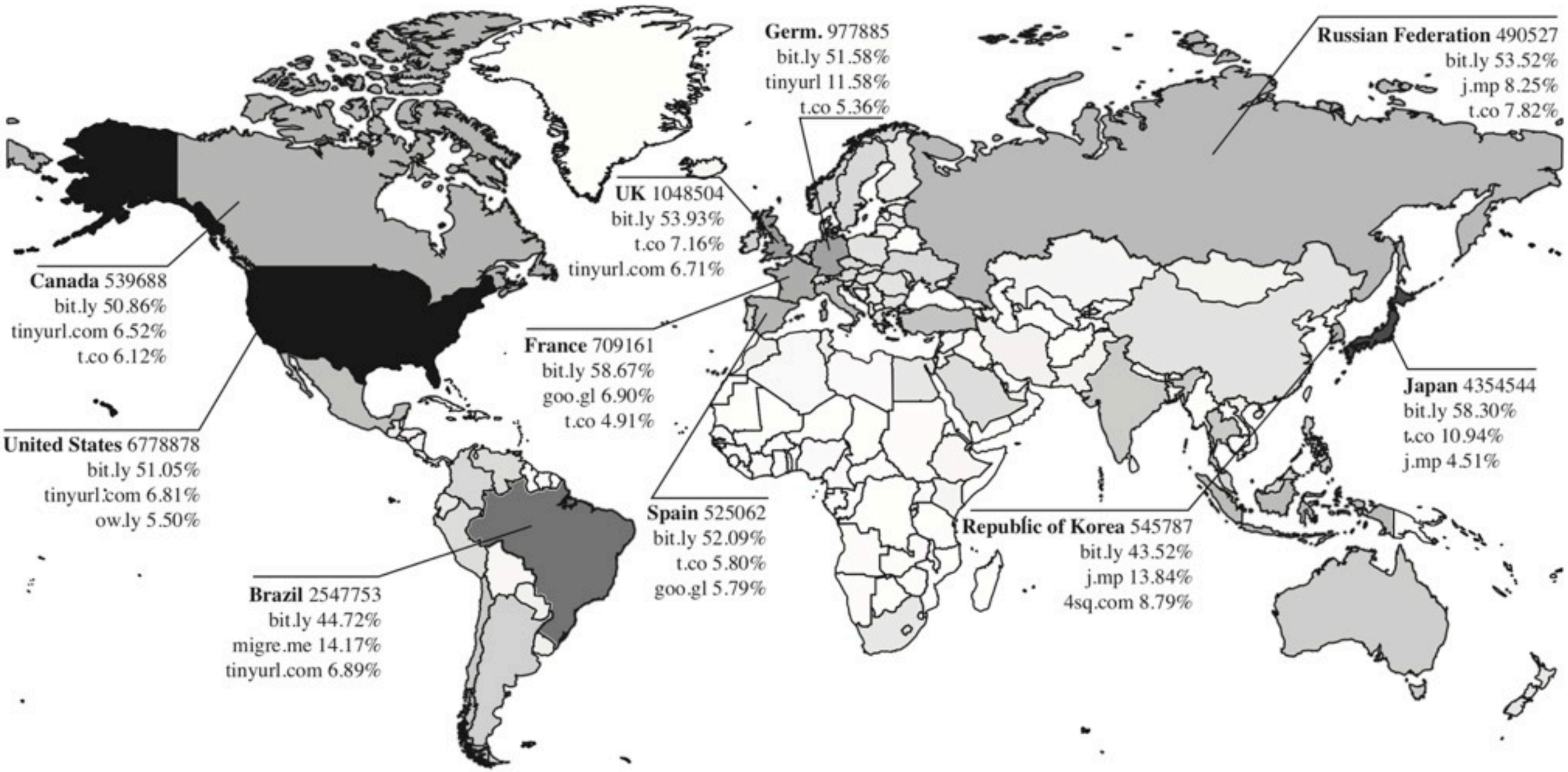
Landing page → Title, Size, Content, ...

Collected data

- Total 7,000 distinct users (estimate from 1,370,277 distinct IPs)
  - about 500 to 1,000 active users per day
  - about 20,000 to 50,000 short URLs sent each day (100,000 peaks)
- 24,953,881 distinct short URLs encountered by users while browsing



Geographical distribution of the collectors (GeoIP)



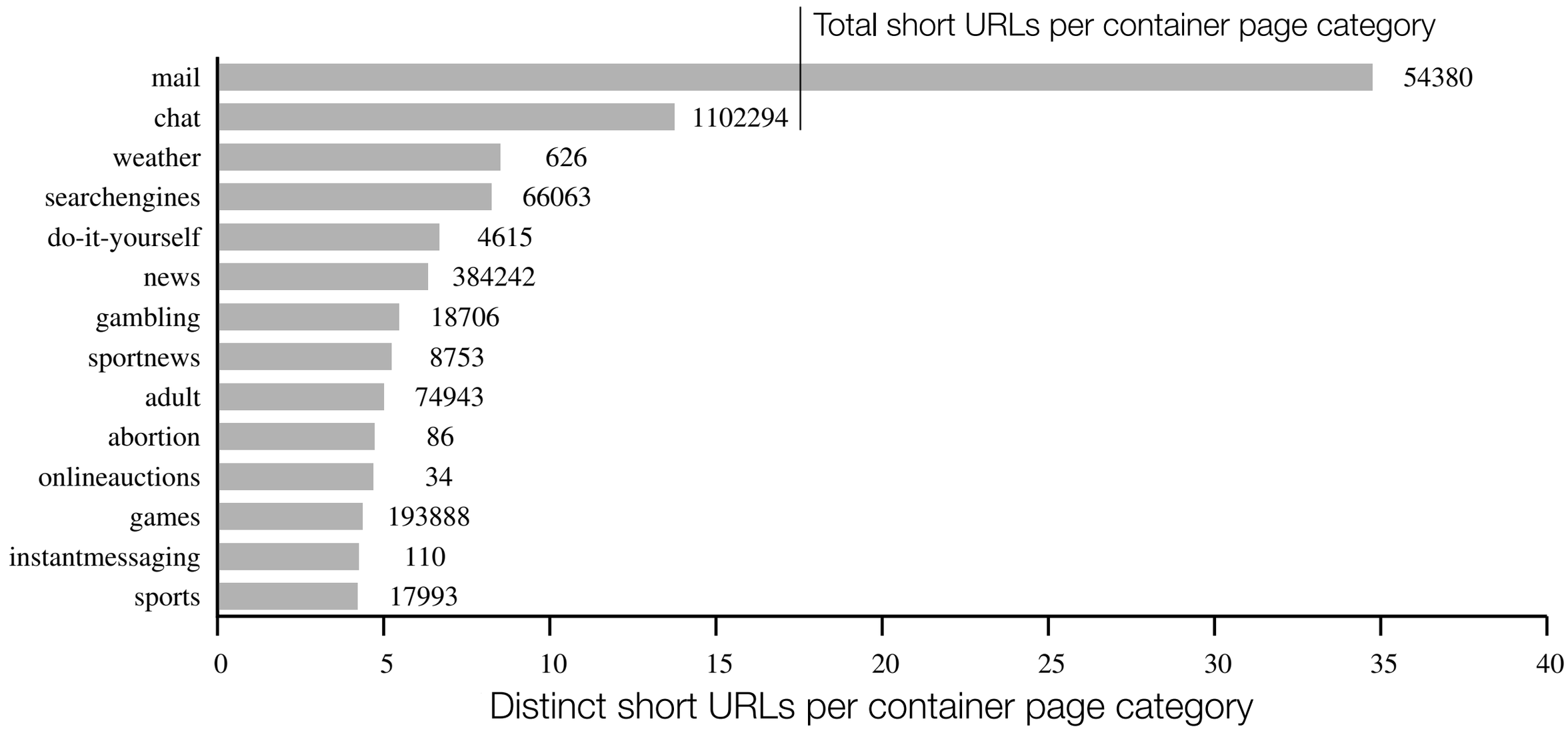
Top services encountered by users while browsing

Distinct URLs		Log entries	
8,179,229	bit.ly	13,407,588	bit.ly
1,047,790	tinyurl.com	2,056,857	tinyurl.com
922,682	t.co	1,658,808	t.co
651,074	ow.ly	1,154,522	ow.ly
607,939	goo.gl	1,045,336	goo.gl
508,969	fb.me	709,444	j.mp
481,398	4sq.com	648,435	is.gd
435,418	tl.gd	618,033	4sq.com
369,960	j.mp	576,815	fb.me
332,118	is.gd	485,221	durl.me

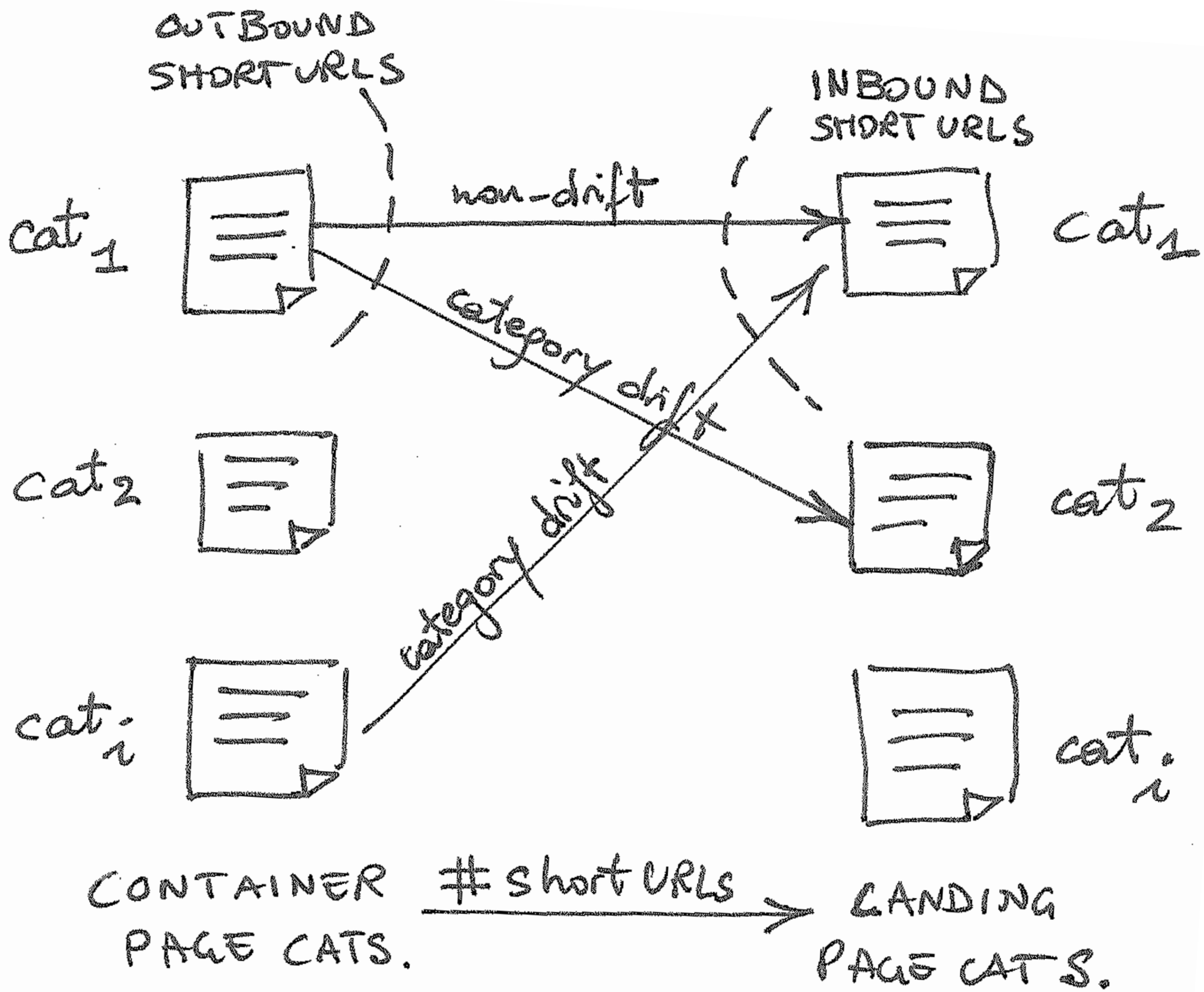
(as of April 2011)

Type of content aliased via short URLs

- We categorize landing pages and container pages
- We use a human-maintained list of categories (DMOZ Open Directory Project)



What happens when users **click** on a short URL?

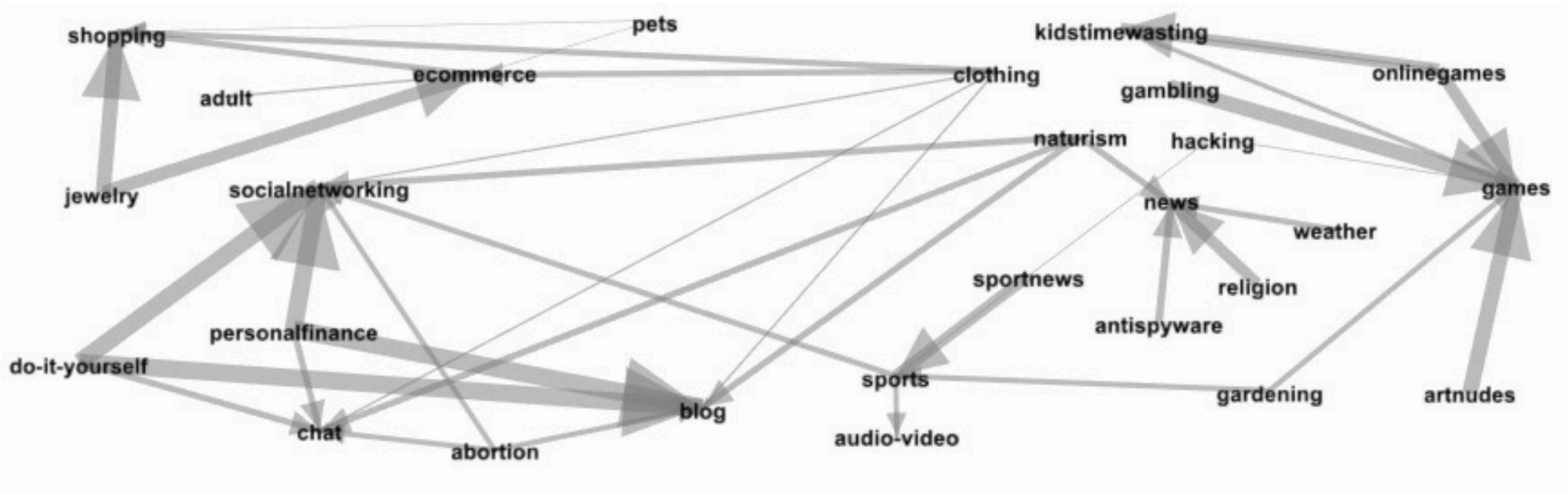

OUTBOUND SHORT URLS
INBOUND SHORT URLS
non-drift
category drift
category drift
cat 1
cat 1
cat 2
cat 2
cat i
cat i
CONTAINER PAGE CATS.
short URLS
LANDING PAGE CATS.

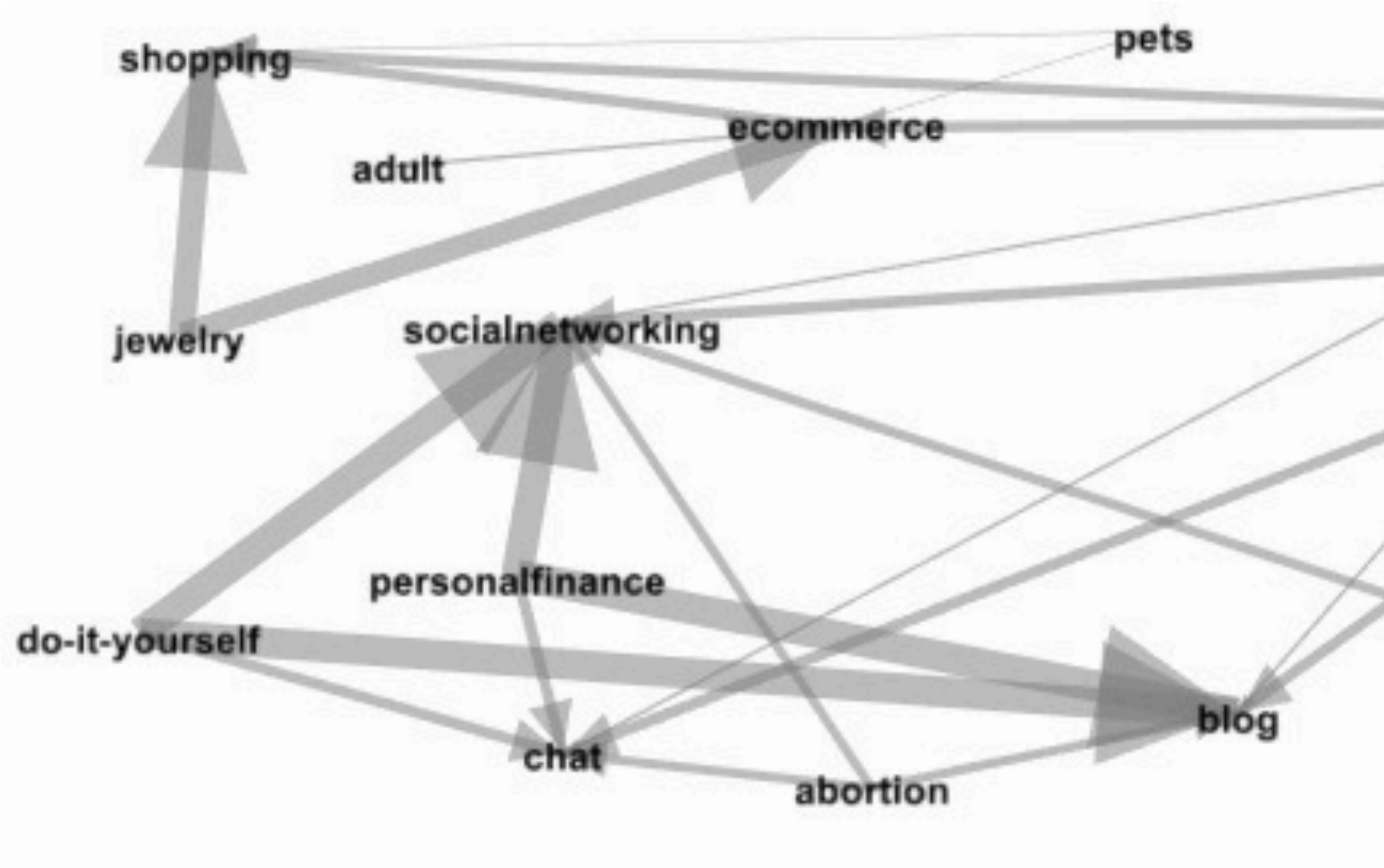
$\rho \to 0$ Many **outbound** short URLs (**aggregators**, e.g., Twitter)

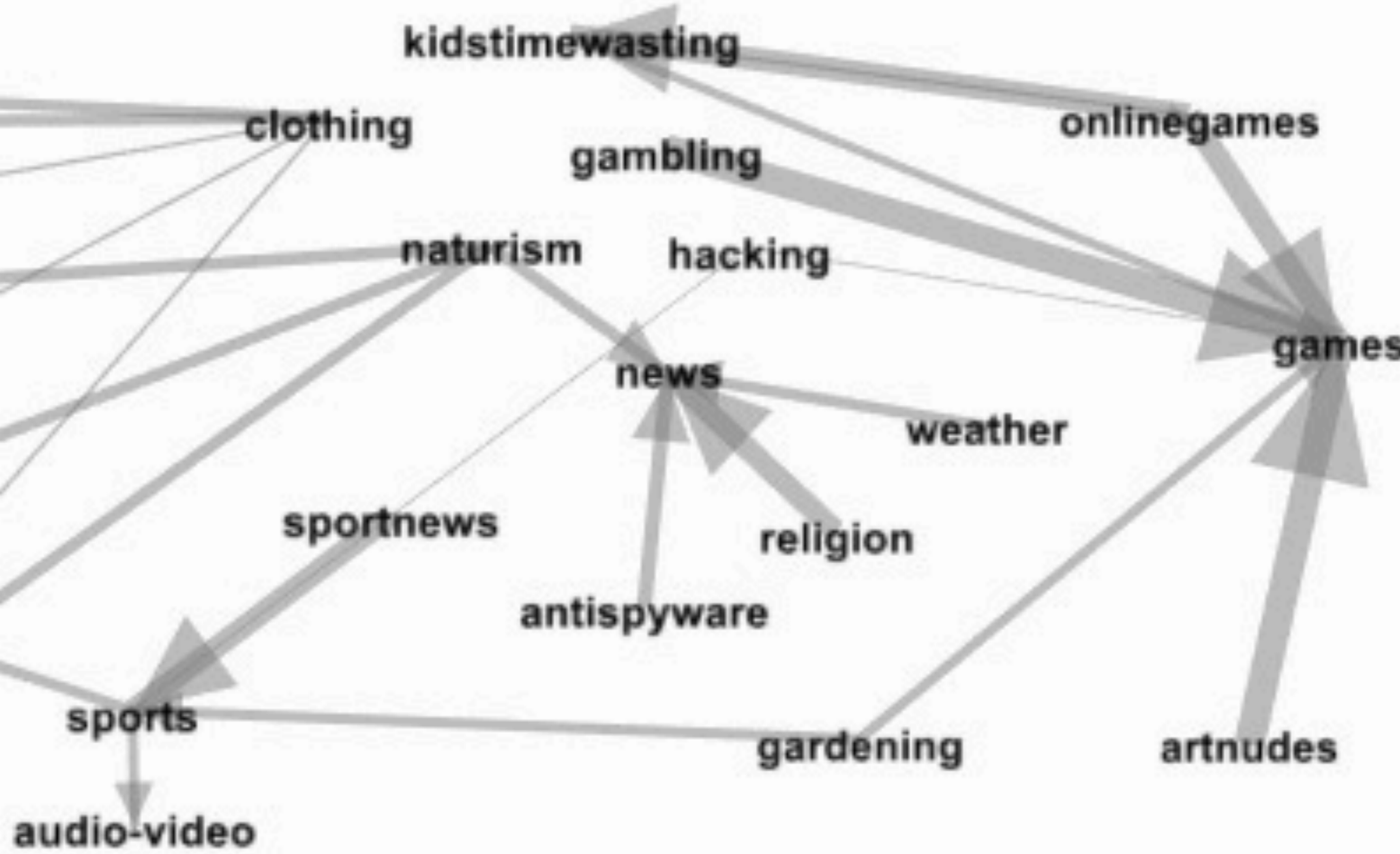
$\rho \to 1$ Many **inbound** short URLs (**landing pages**, e.g., news, blogs)

ρ	Category						
0.00	naturism	0.18	artnudes	0.36	weapons	0.75	shopping
0.01	personalfinance	0.21	antispyware	0.36	cleaning	0.78	games
0.01	do-it-yourself	0.23	drinks	0.37	dating	0.80	news
0.03	pets	0.25	medical	0.39	vacation	0.82	government
0.04	gardening	0.25	weather	0.40	religion	0.88	chat
0.07	clothing	0.30	onlinegames	0.42	culinary	0.90	blog
0.07	mail	0.32	jobsearch	0.45	filehosting	0.91	socialnetworking
0.09	banking	0.33	sportnews	0.52	kidstimewasting	1.00	contraception
0.12	abortion	0.33	gambling	0.55	ecommerce	1.00	childcare
0.12	instantmessaging	0.36	drugs	0.67	adult	1.00	astrology
0.13	jewelry	0.36	searchengines	0.68	audio-video	1.00	cellphones
0.18	hacking	0.36	weapons	0.69	sports	1.00	onlineauctions
						1.00	onlinepayment

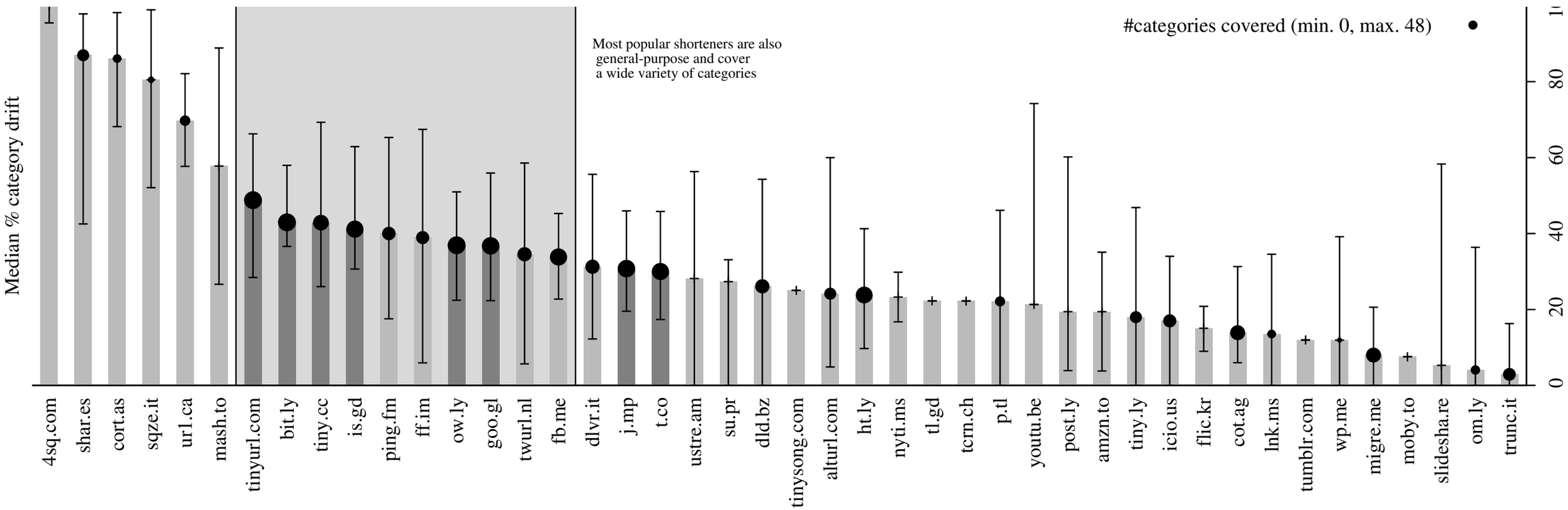
$$\rho = \frac{In(cat)}{In(cat) + Out(cat)}$$

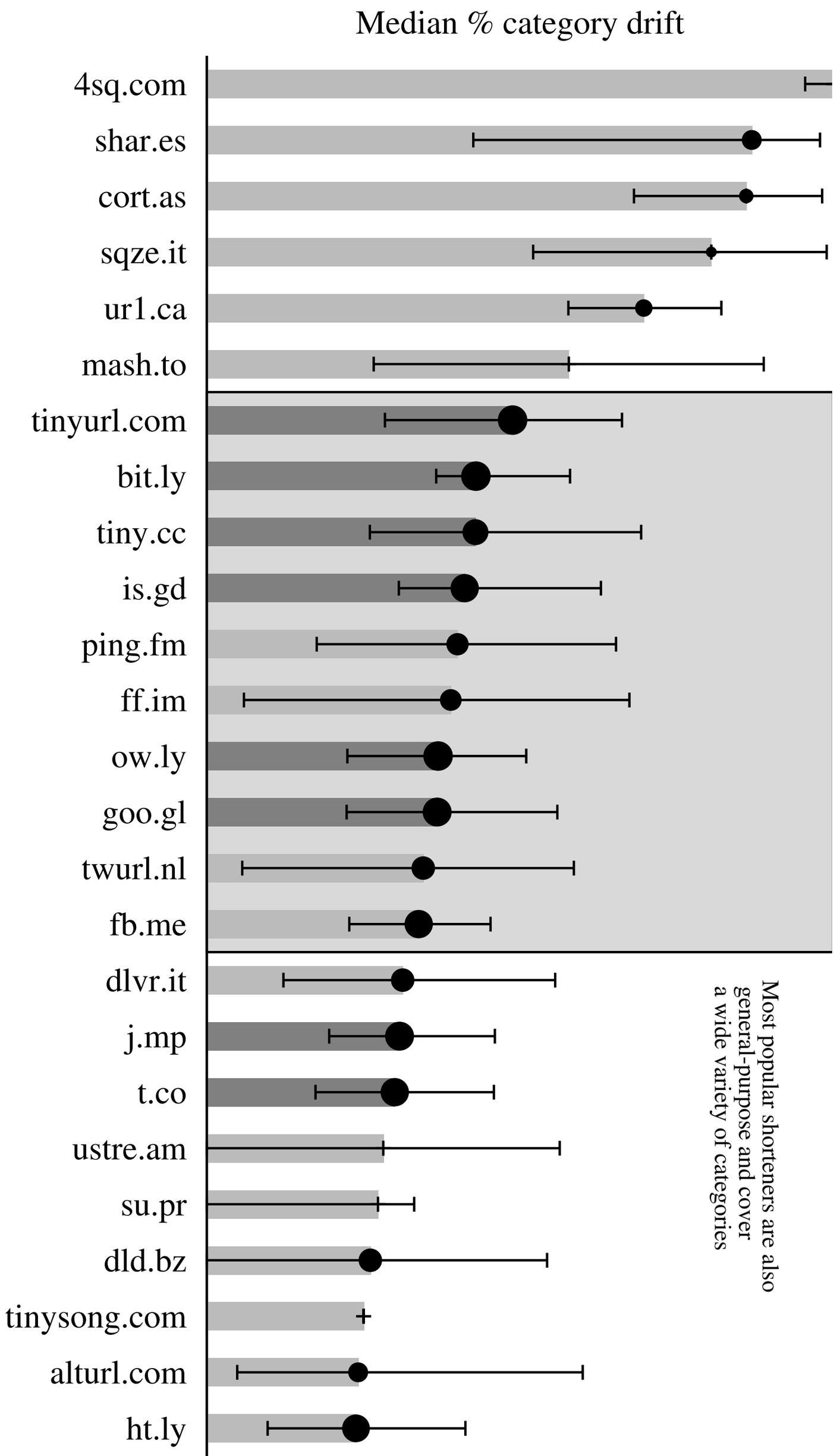

shopping
pets
kidstimewasting
ecommerce
clothing
onlinegames
adult
gambling
naturism
hacking
jewelry
socialnetworking
games
news
weather
sportnews
religion
antispyware
personalfinance
do-it-yourself
sports
gardening
artnudes
blog
chat
audio-video
abortion

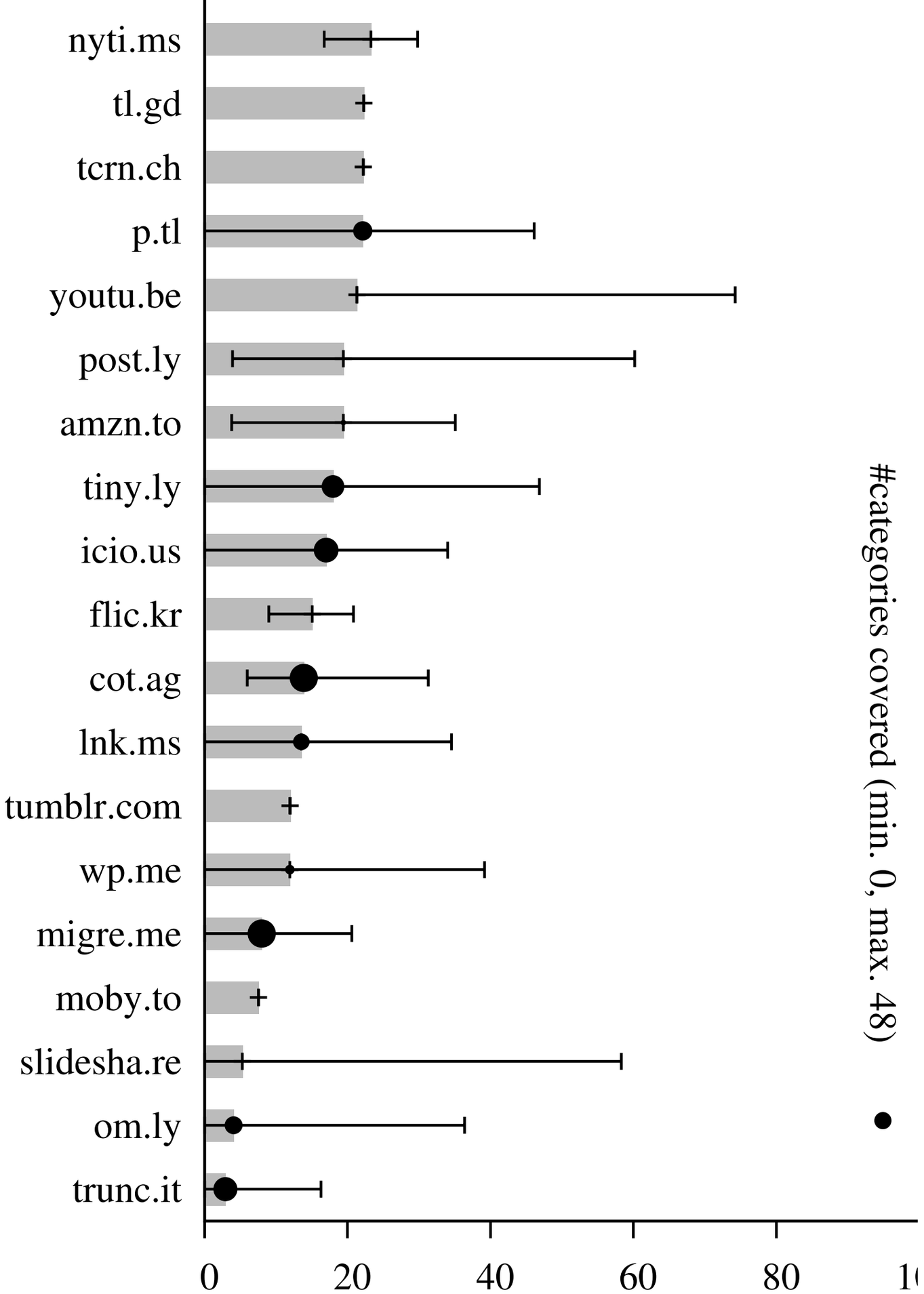

shopping
pets
ecommerce
adult
jewelry
socialnetworking
personalfinance
do-it-yourself
blog
chat
abortion


kidstimewasting
clothing
onlinegames
gambling
naturism
hacking
games
news
weather
sportnews
religion
antispyware
sports
gardening
artnudes
audio-video

Content-specific vs. general-purpose services




Median % category drift
4sq.com
shar.es
cort.as
sqze.it
ur1.ca
mash.to
tinyurl.com
bit.ly
tiny.cc
is.gd
ping.fm
ff.im
ow.ly
goo.gl
twurl.nl
fb.me
dlvr.it
j.mp
t.co
ustre.am
su.pr
dld.bz
tinysong.com
alturl.com
ht.ly
Most popular shorteners are also
general-purpose and cover
a wide variety of categories


nyti.ms
tl.gd
tcrn.ch
p.tl
youtu.be
post.ly
amzn.to
tiny.ly
icio.us
flic.kr
cot.ag
lnk.ms
tumblr.com
wp.me
migre.me
moby.to
slidesha.re
om.ly
trunc.it
0
20
40
60
80
#categories covered (min. 0, max. 48)

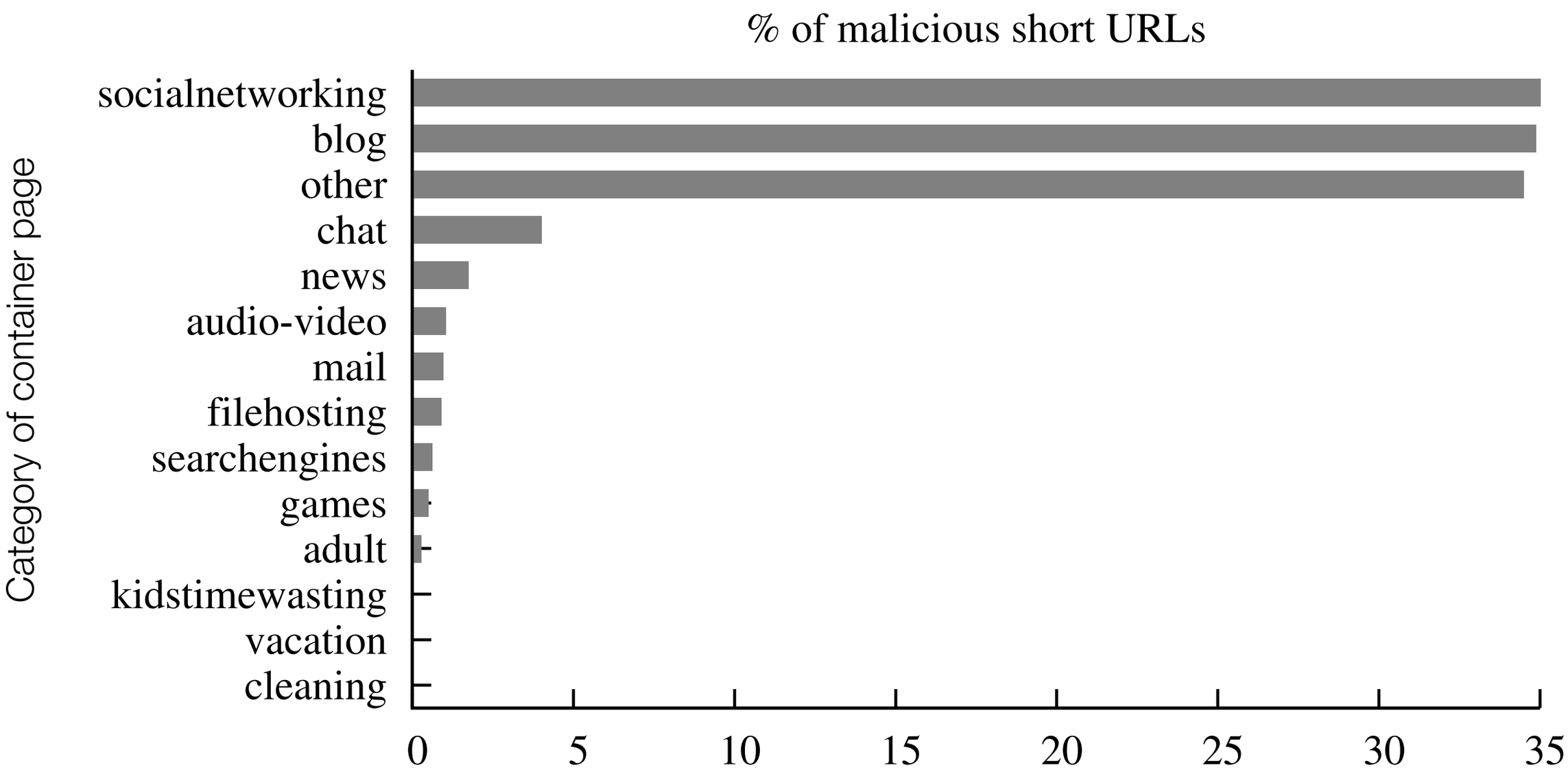
Security aspects related to short URLs

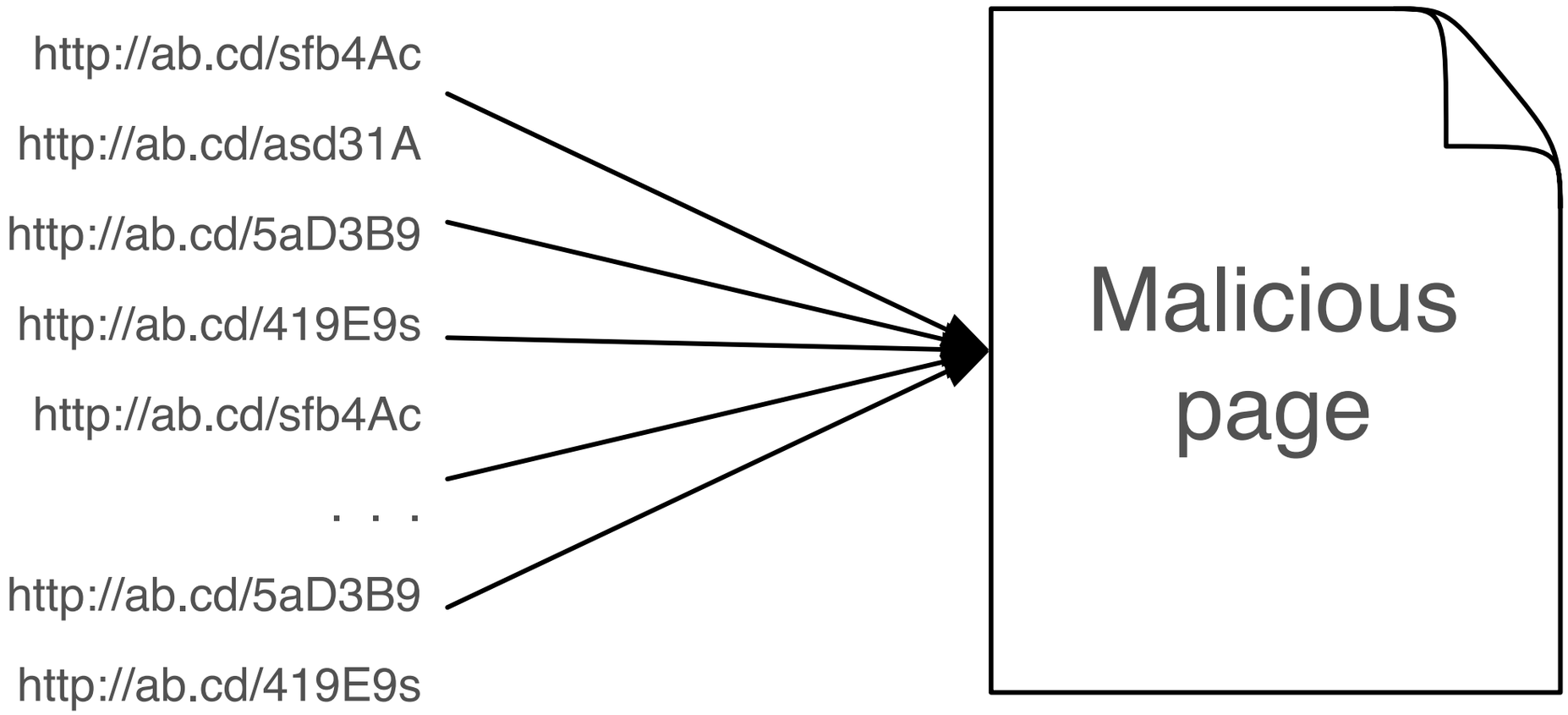
Malicious short URLs encountered by users

Category	Short URLs	Long URLs	Ratio
Phishing	88	79	1.11
Malware	1,161	1,083	1.07
Spam	731	694	1.05

Blacklist	Phishing	Malware	Spam
Spamhaus	-	-	694
Phishtank	61	-	-
Wepawet	-	266	-
Safe Browsing	18	817	-

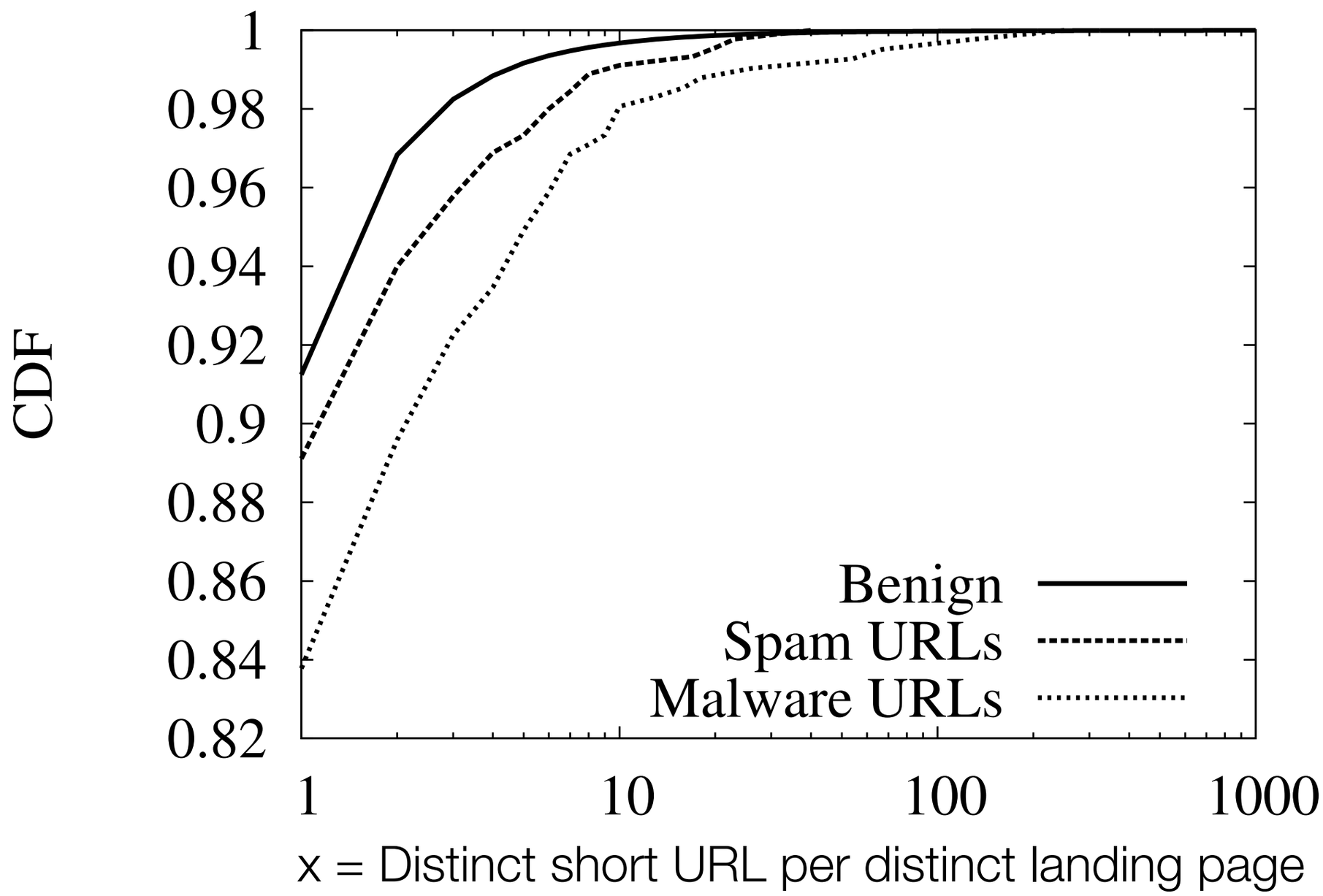
What type of sites contain malicious short URLs?

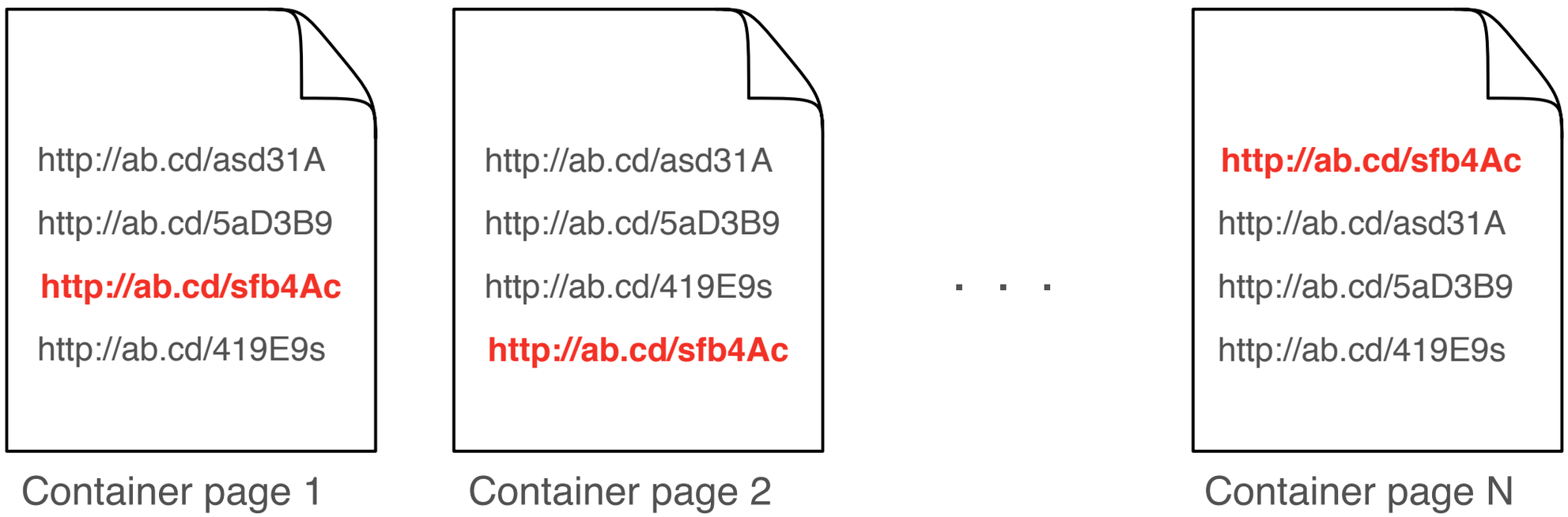



http://ab.cd/sfb4Ac
http://ab.cd/asd31A
http://ab.cd/5aD3B9
http://ab.cd/419E9s
http://ab.cd/sfb4Ac
. . .
http://ab.cd/5aD3B9
http://ab.cd/419E9s
Malicious page

Aliasing of malicious pages using short URLs

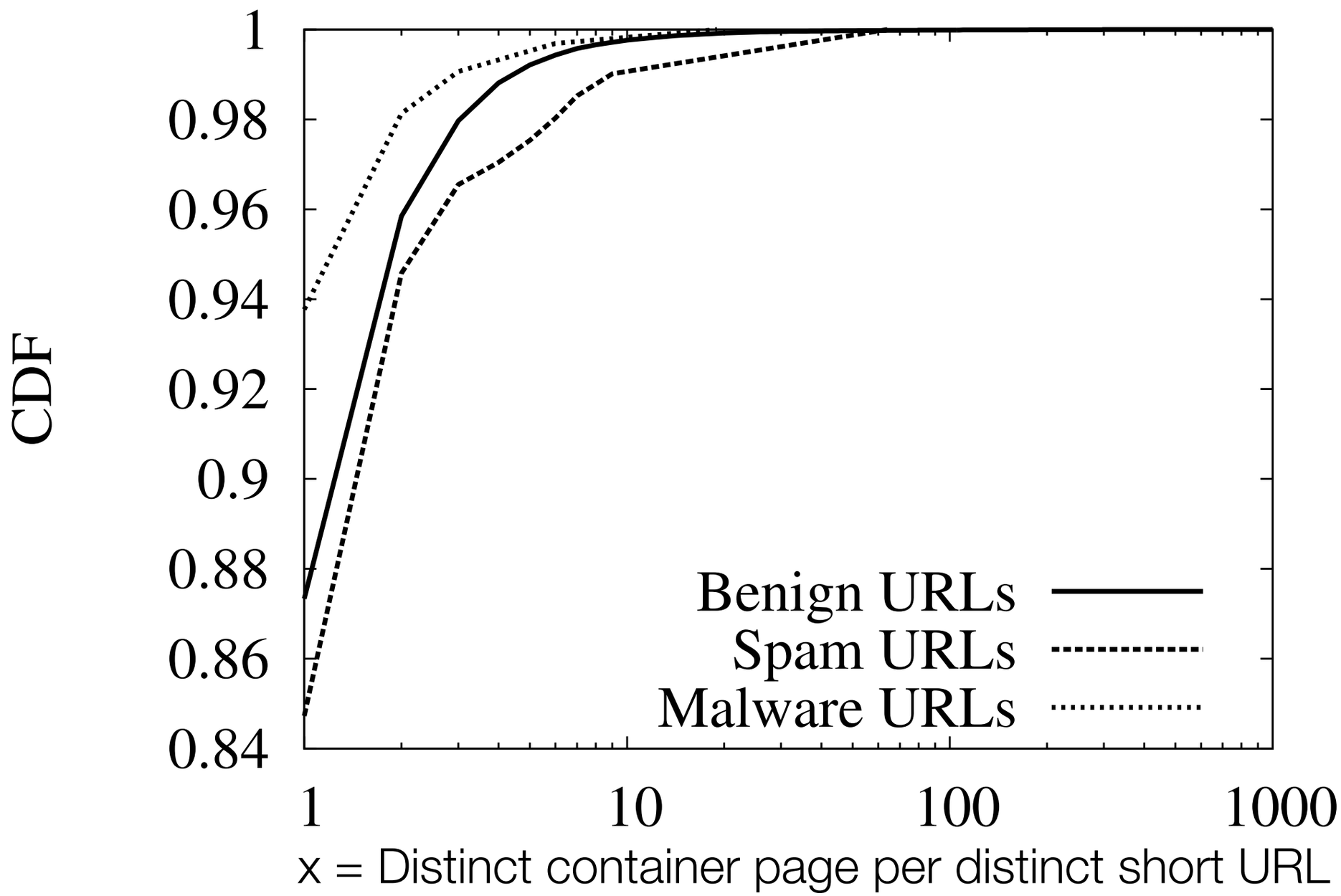
Drive-by and spam landing pages are more aliased than benign ones.




http://ab.cd/asd31A
http://ab.cd/5aD3B9
http://ab.cd/sfb4Ac
http://ab.cd/419E9s
Container page 1
http://ab.cd/asd31A
http://ab.cd/5aD3B9
http://ab.cd/419E9s
http://ab.cd/sfb4Ac
Container page 2
. . .
http://ab.cd/sfb4Ac
http://ab.cd/asd31A
http://ab.cd/5aD3B9
http://ab.cd/419E9s
Container page N

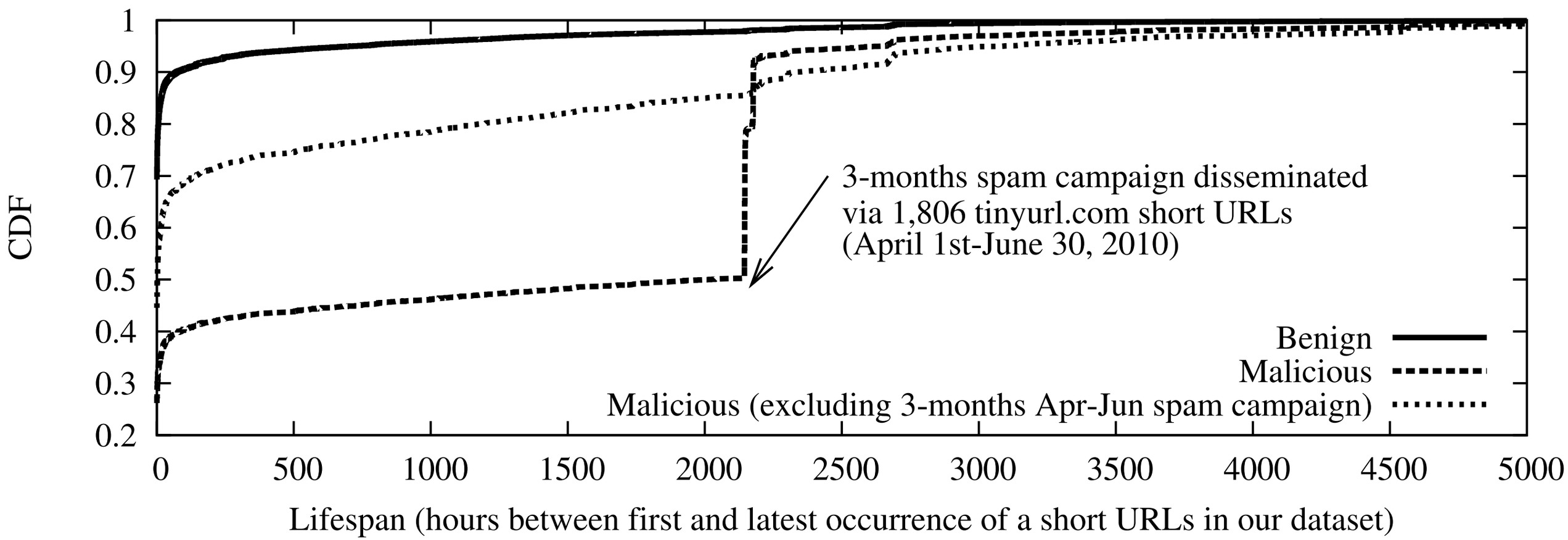
Dissemination of malicious short URLs

Spam short URLs are disseminated on a larger number of container pages.



Lifespan of malicious short URLs

Malicious short URLs typically **survive longer** than benign ones.



Exception: a spam campaign (Storm botnet?) with
1,806 short URLs deleted by tinyurl.com's administrators.

Are shortening services taking countermeasures?

1. Prepare a list of benign and malicious long URLs

2. Shorten them via the top 6 shortening services (e.g., bit.ly, is.gd, tinyurl.com)

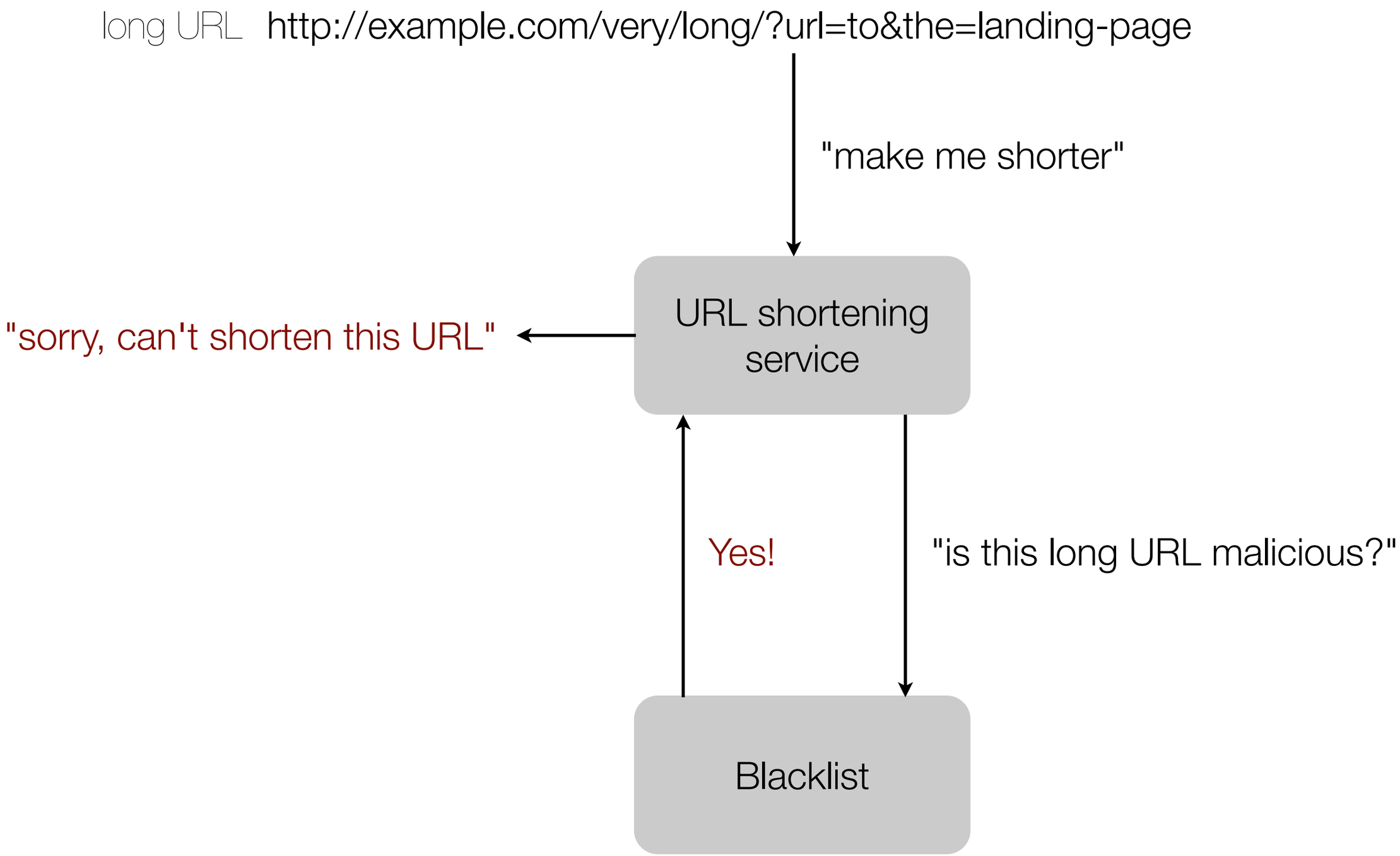
    2.1. Do they **accept** malicious URLs (spam, phishing, drive-by download)?

3. Try to access the malicious shortened URLs

    3.1. Do they **warn** the users when they **resolve** the short URLs?

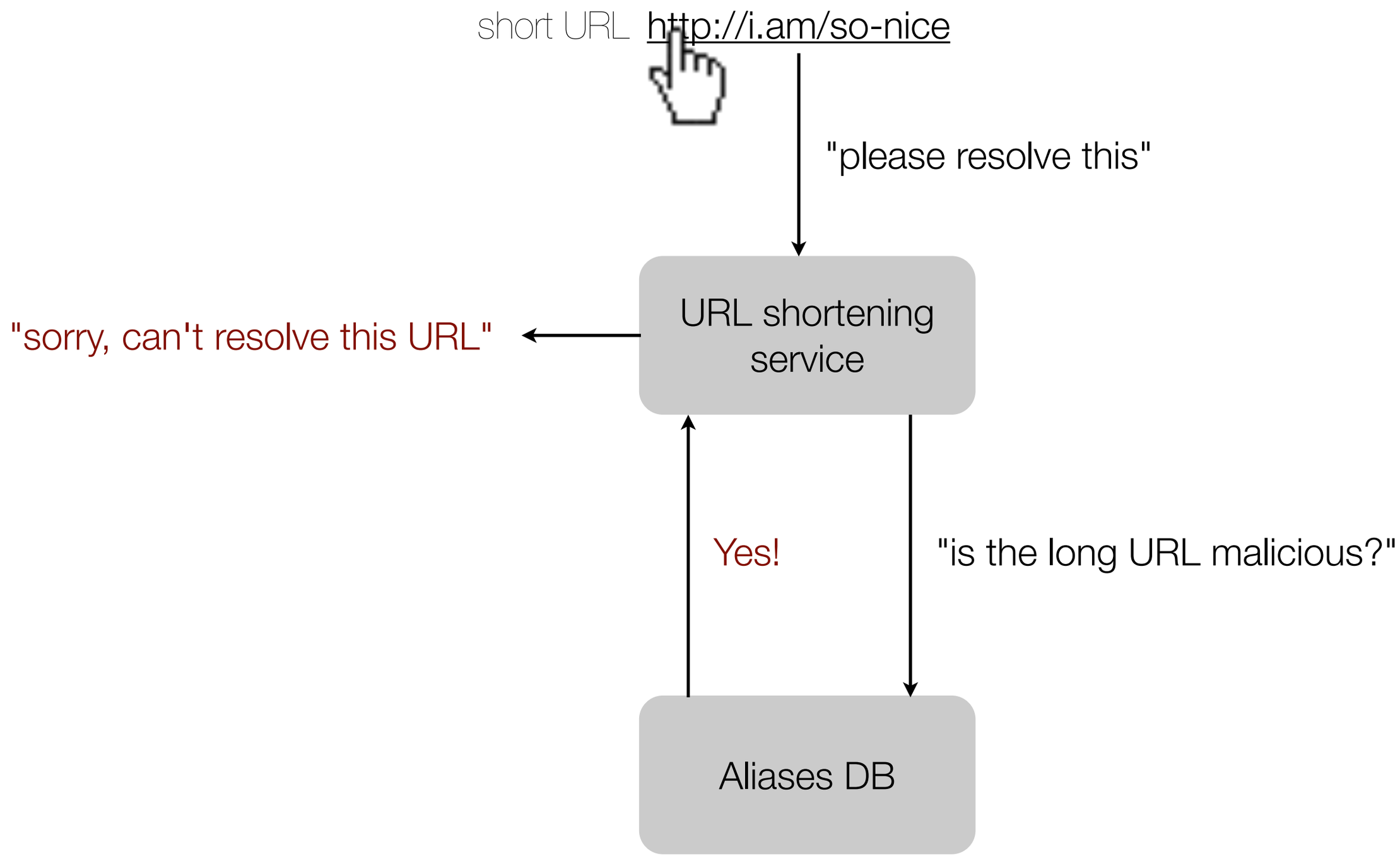
4. Modify the benign long URLs (under our control) and make them malicious

    4.1. Do they **periodically** check their databases of **existing** short URLs?


long URL http://example.com/very/long/?url=to&the=landing-page
"make me shorter"
URL shortening service
"sorry, can't shorten this URL"
Yes!
"is this long URL malicious?"
Blacklist

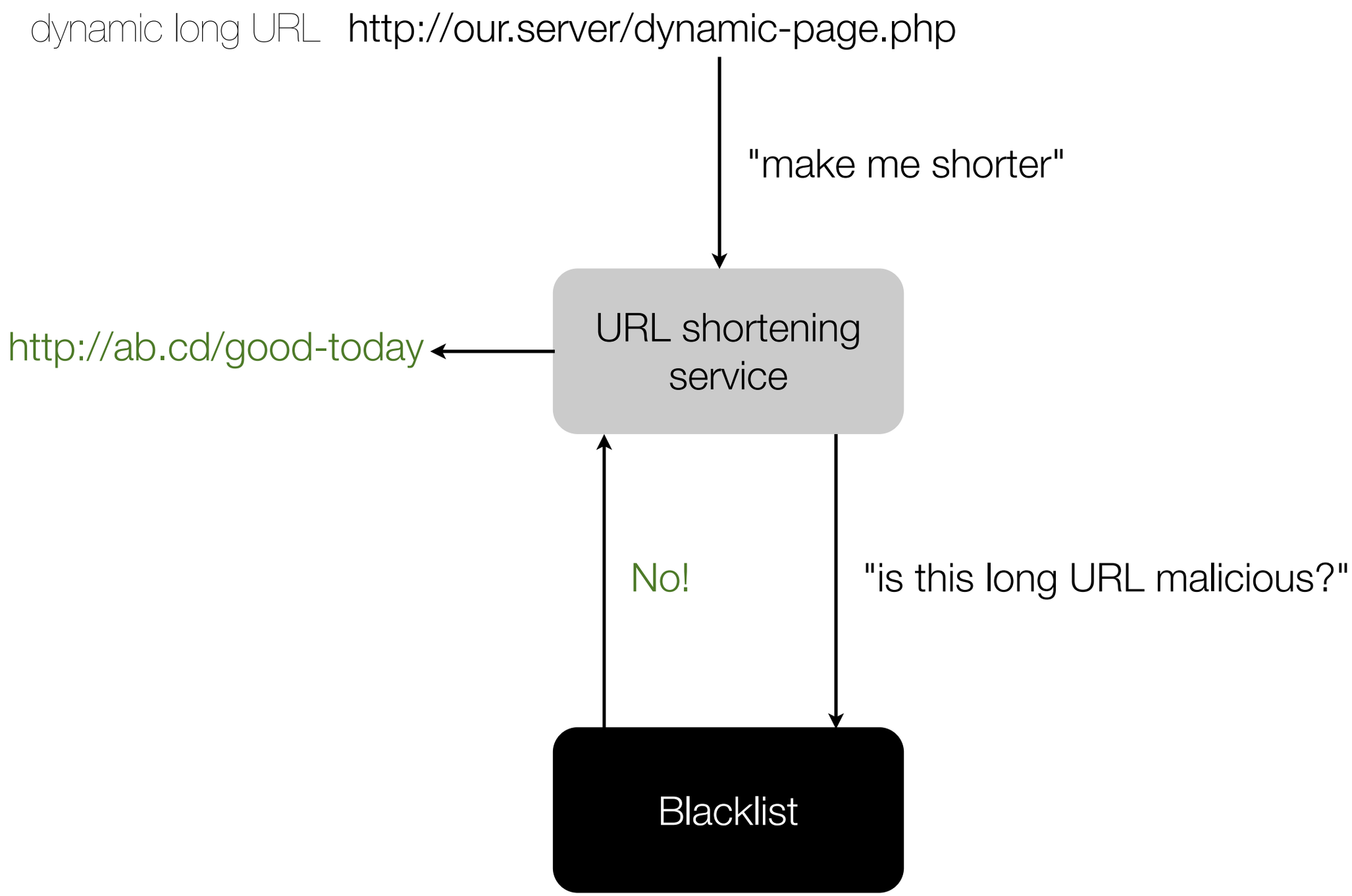
Malicious long URLs accepted by top services

Service	Malware		Phishing		Spam	
	#	%	#	%	#	%
bit.ly	997	99.7	1,000	100.0	1,000	100.0
durl.me	898	89.8	937	93.7	216	21.6
goo.gl	999	99.9	994	99.4	1,000	100.0
is.gd	640	64.0	358	35.8	143	14.3
migre.me	201	20.1	402	40.2	235	23.5
tinyurl.com	997	99.7	996	99.6	998	99.8
Overall	4,732	78.9	4,687	78.1	3,592	59.9

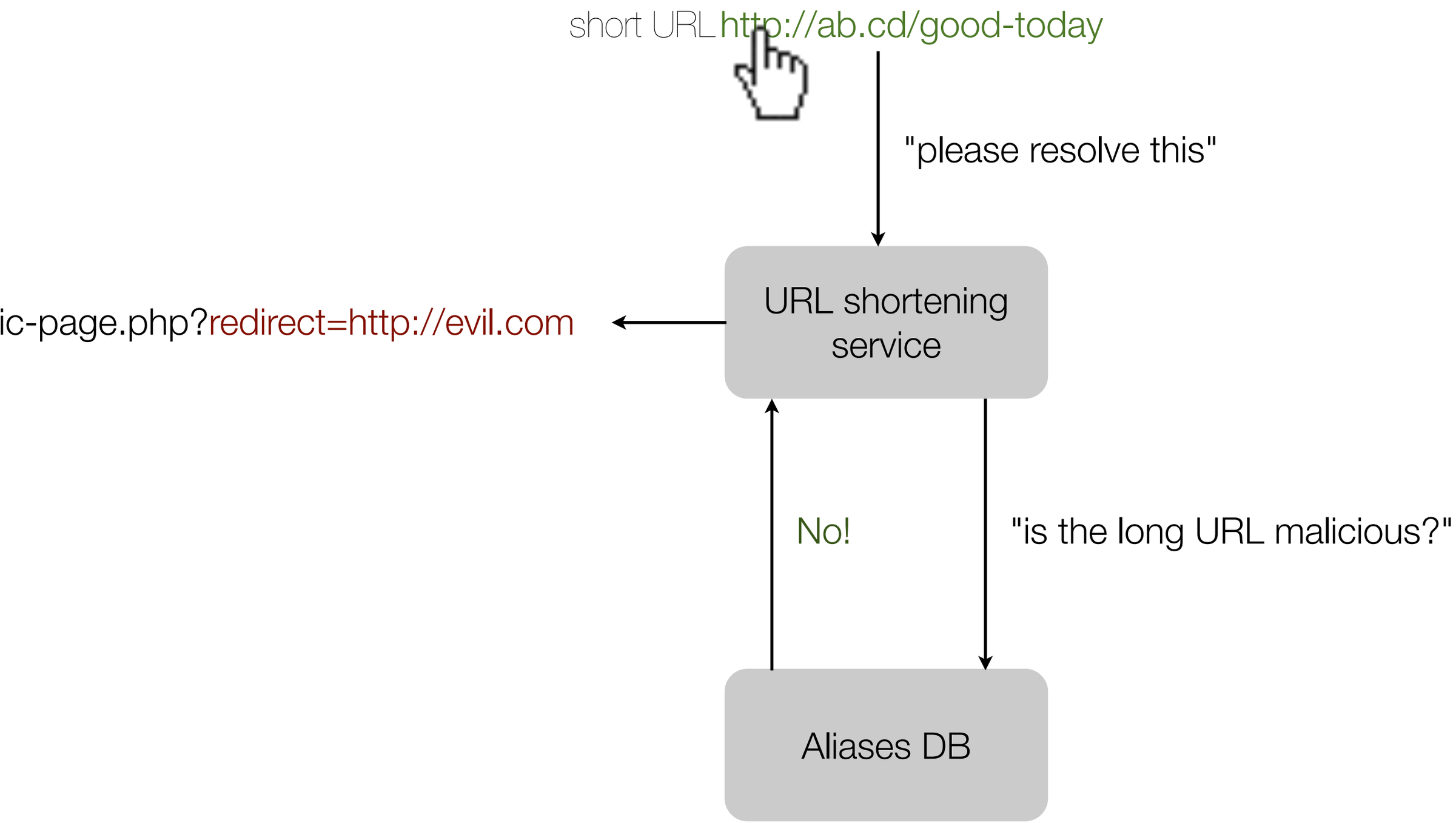

short URL http://i.am/so-nice
"please resolve this"
URL shortening service
"sorry, can't resolve this URL"
Yes!
"is the long URL malicious?"
Aliases DB

Alerting users when accessing bad short URLs

Service	Malware	Phishing	Spam
bit.ly	100.0	97.5	99.9
durl.me	100.0	100.0	100.0
goo.gl	66.4	96.9	78.7
is.gd	43.3	42.9	78.7
migre.me	46.8	40.6	95.7
tinyurl.com	43.5	43.2	77.1
Overall	66.6	70.2	88.4


dynamic long URL
http://our.server/dynamic-page.php
"make me shorter"
URL shortening service
http://ab.cd/good-today
No!
"is this long URL malicious?"
Blacklist

after 24 hours http://our.server/dynamic-page.php?redirect=http://evil.com


short URL http://ab.cd/good-today
"please resolve this"
URL shortening service
ic-page.php?redirect=http://evil.com
No!
"is the long URL malicious?"
Aliases DB

Deferred maliciousness

Threat	Shortened	Blocked	Not Blocked
Malware	162	0%	100%
Phishing	180	0%	100%
Spam	150	0%	100%
Overall	492	0%	100%

Limitations & future work or what we still need to do

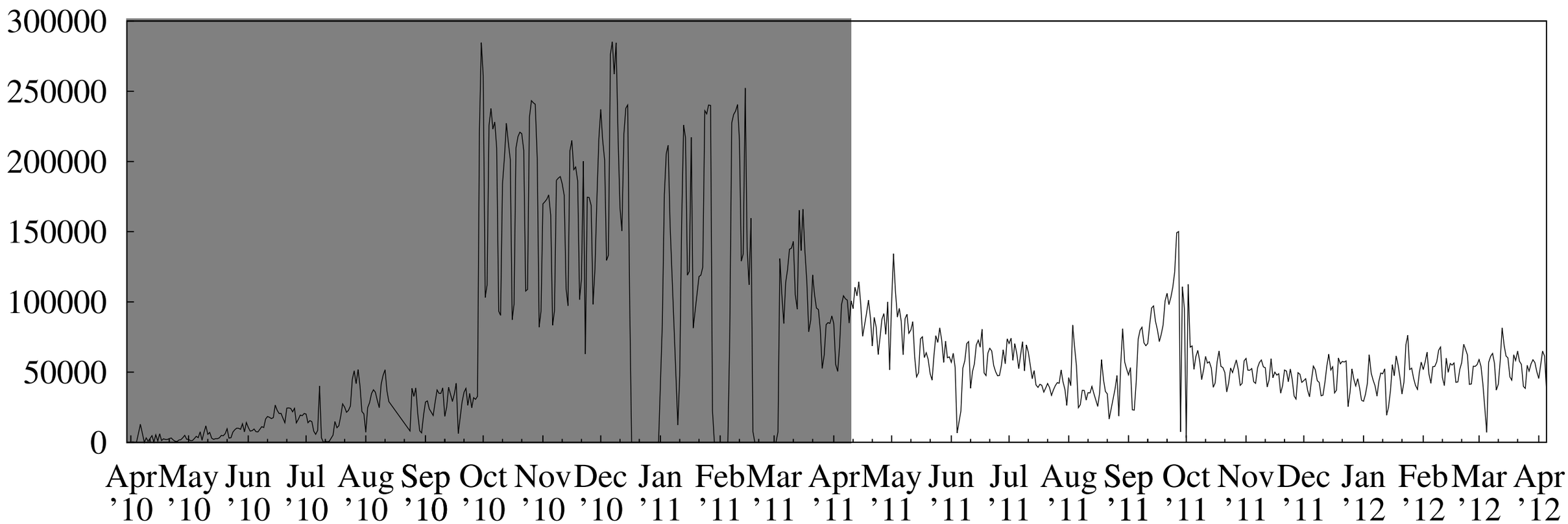
- We collect short URLs only when container pages are visited.
- We track clicks on short URLs, but we collected 42,147 clicks (too early to draw conclusions).
- We have not tracked whether existing, benign short URLs turn into malicious short URLs.

Conclusions: What is the impact on users?

- What do users **use** short URLs for?
  - Share ephemeral resources to user-generated content (e.g., social nets)
- Do users stumble upon short URLs that **often**?
  - Not very often: ~1K over 16M
- Do users **perceive** the maliciousness of a short URL?
  - Not much: almost no one clicked on our "flag as malicious" link. Also confirmed by [Onarlioglu et al., NDSS 2012]
- Do URL shortening services take enough **countermeasures** to protect the users?
  - Some of them use **blacklists** but do **not** proactively check **existing** aliases

We're still collecting short URLs

- 16,075,693 over 24,953,881 analyzed thoroughly
- No big changes in the **new** portion of the dataset



Co-authors

Alessandro Frossi	Politecnico di Milano
Gianluca Stringhini	UC Santa Barbara
Brett Stone-Gross	UC Santa Barbara
Stefano Zanero	Politecnico di Milano
Christopher Kruegel	UC Santa Barbara
Giovanni Vigna	UC Santa Barbara

Questions?

fede@maggi.cc
@phretor