Detection of Hidden Fraudulent URLs within Trusted Sites using Lexical Features

Enrico Sorio Alberto Bartoli Eric Medvet

DIA - University of Trieste

September 5th, 2013

http://machinelearning.inginf.units.it

Table of Contents

MACHINE LEARNING LAB

Security threats to web sites content

Means of attack:

- Fraudulent changes to existing pages
- Fraudulent creation of new *hidden* pages

Attacker's goals:

- defacement
- search/web spam
- malware spreading

Security threats to web sites content

Means of attack:

- Fraudulent changes to existing pages
- Fraudulent creation of new *hidden* pages

Attacker's goals:

- defacement
- search/web spam
- malware spreading

Fraudulent hidden pages

How long to actually detect them?

- 50% of defacements (non-hidden fraudulent pages) are detected later than 1 week after the incident (study says[1])

[1] Bartoli, A.; Davanzo, G.; Medvet, E., *The Reaction Time to Web Site Defacements*, Internet Computing, IEEE , vol.13, no.4, pp.52,58, July-Aug. 2009

Fraudulent hidden pages

How long to actually detect them?

- 50% of defacements (non-hidden fraudulent pages) are detected later than 1 week after the incident (study says[1])
- no studies for hidden pages, but likely longer

[1] Bartoli, A.; Davanzo, G.; Medvet, E., *The Reaction Time to Web Site Defacements*, Internet Computing, IEEE , vol.13, no.4, pp.52,58, July-Aug. 2009

Fraudulent hidden pages

How many of them?

- $>$ 9% Italian PA domains contain fraudulent hidden pages (study says[2])

[2] Sorio, E.; Bartoli, A.; Medvet, E., *A look at hidden web pages in Italian public administrations,* Computational Aspects of Social Networks (CASoN), 2012 Fourth International Conference on , vol., no., pp.291,296, 21-23 Nov. 2012

Fraudulent hidden pages

How many of them?

- $>$ 9% Italian PA domains contain fraudulent hidden pages (study says[2])
- a significan fraction of Zone-H entries are fradulent hidden pages

[2]Sorio, E.; Bartoli, A.; Medvet, E., *A look at hidden web pages in Italian public administrations,* Computational Aspects of Social Networks (CASoN), 2012 Fourth International Conference on , vol., no., pp.291,296, 21-23 Nov. 2012

Fraudulent hidden pages

Are they risky?

- actually used for illicit activities, e.g., illicit drug trade

MACHINE LEARNING LAB

Fraudulent hidden pages

Are they risky?

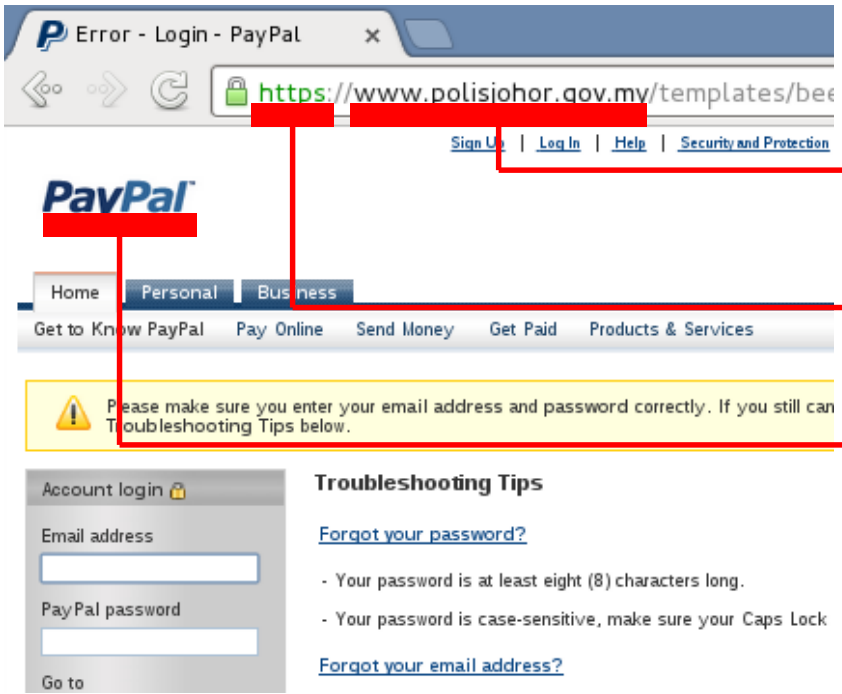
- actually used for illicit activities, e.g., illicit drug trade
- HTTPS does not help: the whole server content is authenticated (by definition) $\rightarrow$ the fraudulent hidden page may live within a trusted site (phishing!)

MACHINE LEARNING LAB

Fraudulent hidden pages

Are they risky?

- actually used for illicit activities, e.g., illicit drug trade
- HTTPS does not help: the whole server content is authenticated (by definition) → the fraudulent hidden page may live within a trusted site (phishing!)



Malaysian police

HTTPS!

PayPal phishing

MACHINE LEARNING LAB

Table of Contents

MACHINE LEARNING LAB

Detection: requirements

Detecting if a page at URL *u* is fraudulent *and* hidden:

- without fetching the page
- without using the *domain* part of *u*
  (~~`http://foo:bar@www.here.com`~~`/path/to/there.htm?p=1#title`)

Detection: requirements

Without using the *domain* part of u

- most existing detectors use the domain as key feature $\Rightarrow$ URLs at domain d are either deemed *all* legitimate or *all* illegitimate
- we exclude the domain and evaluate approach on both legitimate and illegitimate on the *same* domain

MACHINE LEARNING LAB

Method variants

Three variants:

- lexical features
- lexical features and headers
- lexical features, headers and age

Method variants

Three variants:

- lexical features
- lexical features and headers
- lexical features, headers and age

few features, low complexity

↓

many features, high complexity

Variants complexity

- Lexical $\rightarrow$ no HTTP requests
- Lexical+headers $\rightarrow$ one HEAD request
- Lexical+headers+age $\rightarrow$ two HEAD requests

Variants complexity

- Lexical → no HTTP requests can work offline!
- Lexical+headers → one HEAD request
- Lexical+headers+age → two HEAD requests

Method phases

Two phases:

1. calibration, using a *training set* U of URLs and corresponding labels L
2. actual classification

Lexical

Calibration:

1. remove all up to domain from URLs (~~http://foo:bar@www.here.com~~/path/to/there.htm?p=1#title)
2. compute unigrams ($u_i \in U \rightarrow f_i \in \mathbb{R}^n$)
3. train a SVM on $\langle f_i, l_i \rangle$

Lexical

Calibration:

1. remove all up to domain from URLs (~~http://foo:bar@www.here.com~~/path/to/there.htm?p=1#title)
2. compute unigrams ($u_i \in U \rightarrow f_i \in \mathbb{R}^n$)
3. train a SVM on $\langle f_i, l_i \rangle$

	Lexical				
URL	a	b	c	...	l
/p/attc.txt?ac	2	0	2	...	+
/bb/user/car/38	1	2	1	...	−
/xXx.htm	0	0	0	...	+

Lexical

Classification of u:

1. remove all up to domain from u
2. compute unigrams ($u \rightarrow f$)
3. apply SVM to f

MACHINE LEARNING LAB

Lexical+headers

As above, but with more features:

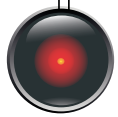
1. obtain by HTTP HEAD header values of
   - `Server` (only major+minor version)
   - `X-Powered-By` (only major+minor version)
   - `Content-Type`
   - `Content-Length`
2. map (all but `Content-Length`) as binary features

Lexical+headers

As above, but with more features:

1. obtain by HTTP HEAD header values of
   - `Server` (only major+minor version)
   - `X-Powered-By` (only major+minor version)
   - `Content-Type`
   - `Content-Length`
2. map (all but `Content-Length`) as binary features

	Lexical				Server				Content-Type				...	
URL	a	b	c	...	Apache/2.2	Apache/2.4	IIS/7.2	...	text/plain	text/html	text/css	...	...	/
c/p/attc.txt?ac	2	0	2	...	0	1	0	...	1	0	0	...	...	+
/bb/user/car/38	1	2	1	...	1	0	0	...	0	1	0	...	...	−
/xXx.htm	0	0	0	...	1	0	0	...	0	1	0	...	...	+



MACHINE LEARNING LAB

Lexical+headers+age

As above, but with one more feature (the *relative age*):

1. obtain the `Last-Modified` header value of u
2. obtain the `Last-Modified` header value of *home page* of u (`http://www.carexperts.it/bb/user/car/38` $\rightarrow$ `http://www.carexperts.it`)
3. use the difference as relative age

Table of Contents

MACHINE LEARNING LAB

Dataset

Evaluating a detector of hidden fraudulent pages

- no public dataset available
- we built one (and made publicly available[3])

[3] http://machinelearning.inginf.units.it/data-and-tools/hidden-fraudulent-urls-dataset

Positives and negatives

Positive examples (+):

- hidden fraudulent pages

Negative examples (−):

- “normal” pages on “normal” sites
- “normal” pages on attacked sites

Positive examples (+)

Hidden fraudulent pages:

- validated phishing attacks from Phishtank[4] ($U_+^P \rightarrow 6564$)
- validated defacement attacks from Zone-H[5] ($U_+^D \rightarrow 2144$)

[4] http://www.phishtank.com

[5] http://www.zone-h.org

Positive examples (+)

Hidden fraudulent pages:

- validated phishing attacks from Phishtank[4] ($U_+^P \rightarrow 6564$)
- validated defacement attacks from Zone-H[5] ($U_+^D \rightarrow 2144$)

Hidden $\rightarrow$ not reachable while crawling the site within 3rd level

[4] http://www.phishtank.com

[5] http://www.zone-h.org

MACHINE LEARNING LAB

Negative examples (−)

"Normal" pages on "normal" sites:

- 20 among 500 Alexa top sites, crawled up to 10th level ($U_- \rightarrow 3713$)

"Normal" pages on attacked sites:

- sites of positives phishing attacks, crawled up to 3rd level ($U_-^P \rightarrow 78388$)
- sites of positives defacements attacks, crawled up to 3rd level ($U_-^D \rightarrow 94370$)

MACHINE LEARNING LAB

Experiments procedure

Two suites (phishing, defacements)

- balanced training set and testing set
- 90% training, 10% testing
- 5 repetitions

Results

Phishing:

Method	Accuracy (%)		FNR on U_+^P (%)		FPR on U_-^P (%)		FPR on U_- (%)	
	Avg.	Dev. Std.	Avg.	Dev. Std.	Avg.	Dev. Std.	Avg.	Dev. Std.
Lexical	92.50	0.62	9.80	0.85	5.20	0.84	4.93	1.21
Lexical+headers	95.34	0.48	4.93	1.05	4.38	0.62	0.79	0.70
Lexical+headers+age	95.57	0.37	4.87	1.06	3.98	0.52	0.73	0.73

Defacements:

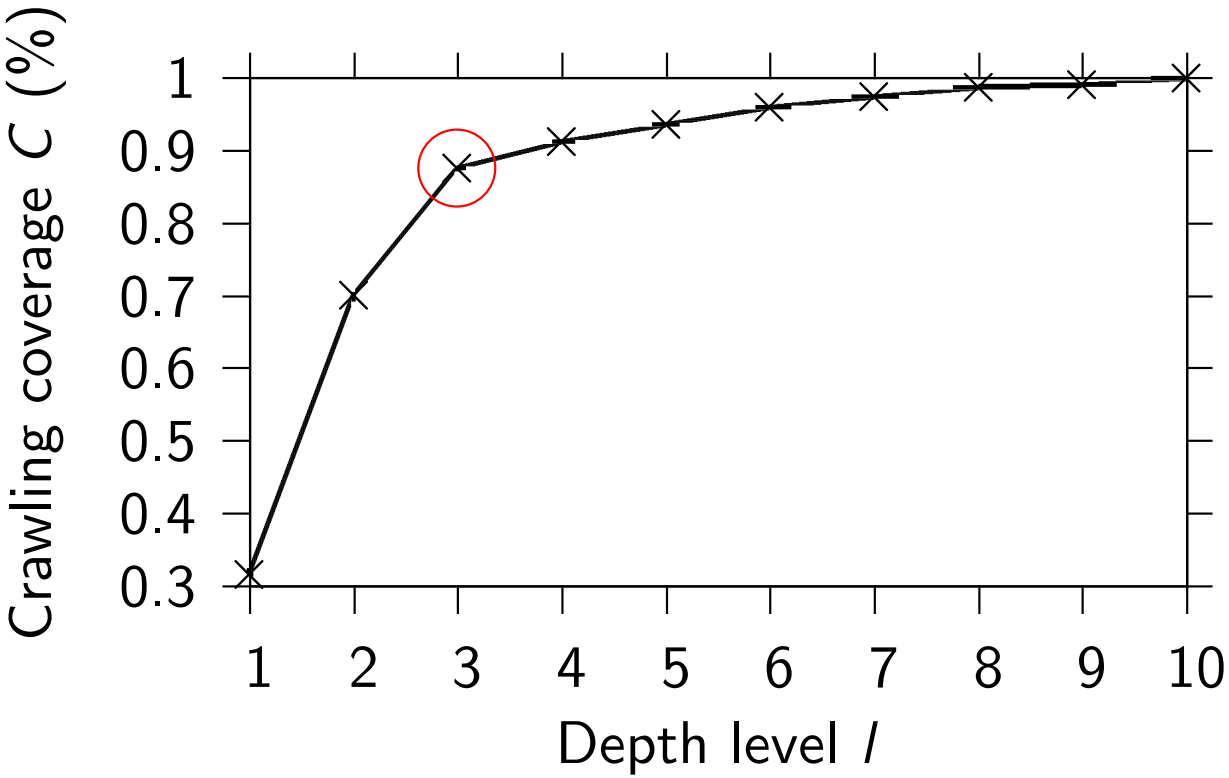
Method	Accuracy (%)		FNR on U_+^D (%)		FPR on U_-^D (%)		FPR on U_- (%)	
	Avg.	Dev. Std.	Avg.	Dev. Std.	Avg.	Dev. Std.	Avg.	Dev. Std.
Lexical	98.37	0.68	0.93	0.74	2.32	0.87	2.41	1.92
Lexical+headers	99.35	0.3	0.37	0.61	0.93	0.57	0.93	0.65
Lexical+headers+age	99.26	0.3	0.47	0.66	1.02	0.61	0.93	0.65

Thanks!

MACHINE LEARNING LAB

Hidden flag: validation of the assumption

"Hidden → page not reachable within 3rd level of crawling"



- 20 among 500 Alexa top sites
- 88% of non-hidden pages are reachable withn 3rd level

Positives and negatives

Positive examples (+):

- hidden fraudulent pages

Negative examples (−):

- non-hidden fraudulent pages
- non-hidden non-fraudulent pages

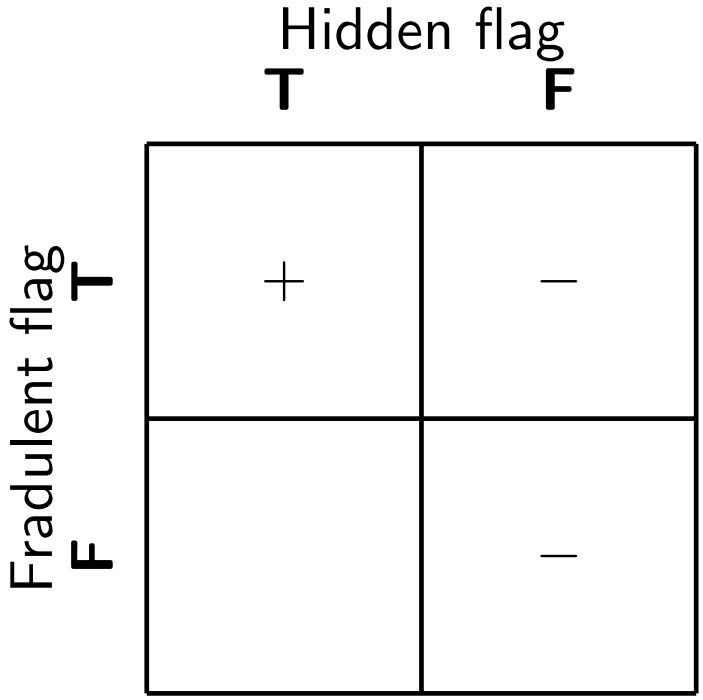
Positives and negatives

Positive examples (+):

- hidden fraudulent pages

Negative examples (−):

- non-hidden fraudulent pages
- non-hidden non-fraudulent pages



Fradulent flag

T $\rightarrow$ fraudulent pages

- validated phishing attacks from Phishtank[6]
- validated defacement attacks from Zone-H[7]

F $\rightarrow$ legitimate pages

- 20 among 500 Alexa top sites, crawled up to 10th level
- Phishtank phished sites, crawled up to 3rd level
- Zone-H defaced attacks sites, crawled up to 3rd level

[6] `http://www.phishtank.com`
[7] `http://www.zone-h.org`

MACHINE LEARNING LAB

Hidden flag

T $\rightarrow$ hidden pages

- not reachable while crawling the site

F $\rightarrow$ non-hidden pages

- reachable while crawling the site

Hidden flag

T $\rightarrow$ hidden pages

- not reachable while crawling the site within 3rd level

F $\rightarrow$ non-hidden pages

- reachable while crawling the site within 3rd level

MACHINE LEARNING LAB

Summary

U^P_+, hidden phishing $\rightarrow$ 6564
U^P_-, non-hidden phishing $\rightarrow$ 78388
U^D_+, hidden defacements $\rightarrow$ 2144
U^D_-, non-hidden defacements $\rightarrow$ 94370
U_-, Alexa $\rightarrow$ 3713

MACHINE LEARNING LAB

Summary

U_+^P, hidden phishing $\rightarrow$ 6564
U_-^P, non-hidden phishing $\rightarrow$ 78388
U_+^D, hidden defacements $\rightarrow$ 2144
U_-^D, non-hidden defacements $\rightarrow$ 94370
U_-, Alexa $\rightarrow$ 3713