Safe Browsing: to Track, Censor and Protect

Thomas Gerbet, Amrit Kumar, Cédric Lauradoux

26 mars 2015

Google Safe Browsing

- **Mise en service** en 2008 pour les navigateurs :
  - Google Chrome
  - Mozilla Firefox
  - Apple Safari
  - Opera
- **Impact :** plus d'un milliard d'utilisateurs selon Google
- **Objectifs :** prévenir les utilisateurs d'atteindre des sites de
  - *phishing*
  - *malwares*
- **Approche :** blacklist
- API compatible avec `C#`, `Python` et `PHP`
- Copier par Yandex (et Baidu ?).

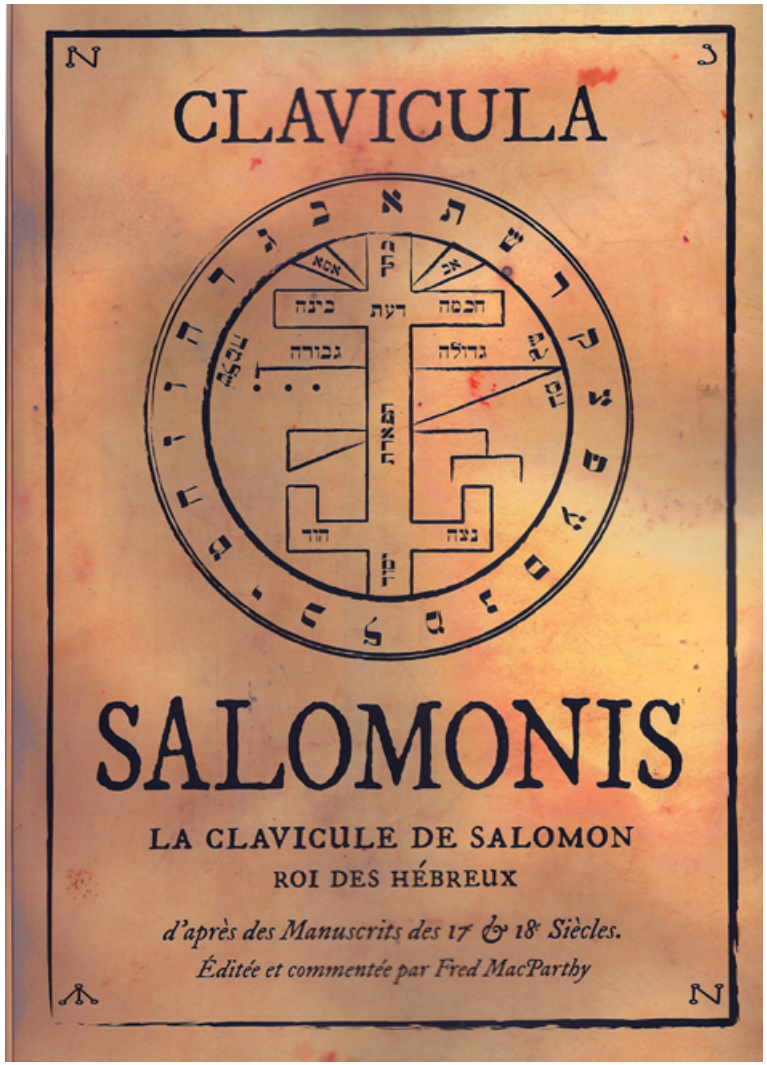
Safe Browsing Lookup API

- GOOGLE **crawle** le web à la recherche des sites de phishing ou de malwares pour **maintenir une blacklist** sur ces serveurs.
- **Utilisation la plus simple :** on demande aux serveurs de GOOGLE si un site est malicieux avec un simple `HTTP GET`.

```
https://sb-ssl.google.com/safebrowsing/api/lookup?
```

- **Problèmes :**
  - passage a l'échelle mauvais
  - problème de protection de la vie privée

La première blacklist de l'histoire



72 Démons de Salomon

Agares

Aim

Alloces

Amdusias

Amon

Amy

Andras

Andrealphus

Andromalius

⋮

Comment reconnaître un démon ?

- **Problème :** tout le monde ne peut pas disposer de la clavicule de Salomon dans sa poche. **Comment faire alors ?**

- **Solution :** faire de la **compression avec perte**.

Ag
Ai
Al
Am
An
⋮

- On passe de 72 noms à 50 préfixes (30% de compression).
- On passe de 518 caractères à 150 (70% de compression).

Les faux positifs

- `Hollande` $\rightarrow$ `Ho` n'est pas dans le dictionnaire. Donc `Hollande` n'est pas un démon.

- `Vals` $\rightarrow$ `Va` **est dans le dictionnaire**. Pourtant `Vals` n'est pas dans la liste complète. C'est un faux positif !

- Si une valeur est dans la liste raccourcie, il faut **vraiment ouvrir la Clavicule de Salomon** pour avoir la liste de tous les démons. Pour `Va`, on aurait : `Valefar`, `Vapula` et `Vassago`.

- **La solution est intéressante si on a peu de faux positifs.**

- Si tout est clair, il est temps de revenir a Safe Browsing.

Google Safe Browsing API v3

- La vérification en local est faites dans ces fichiers :

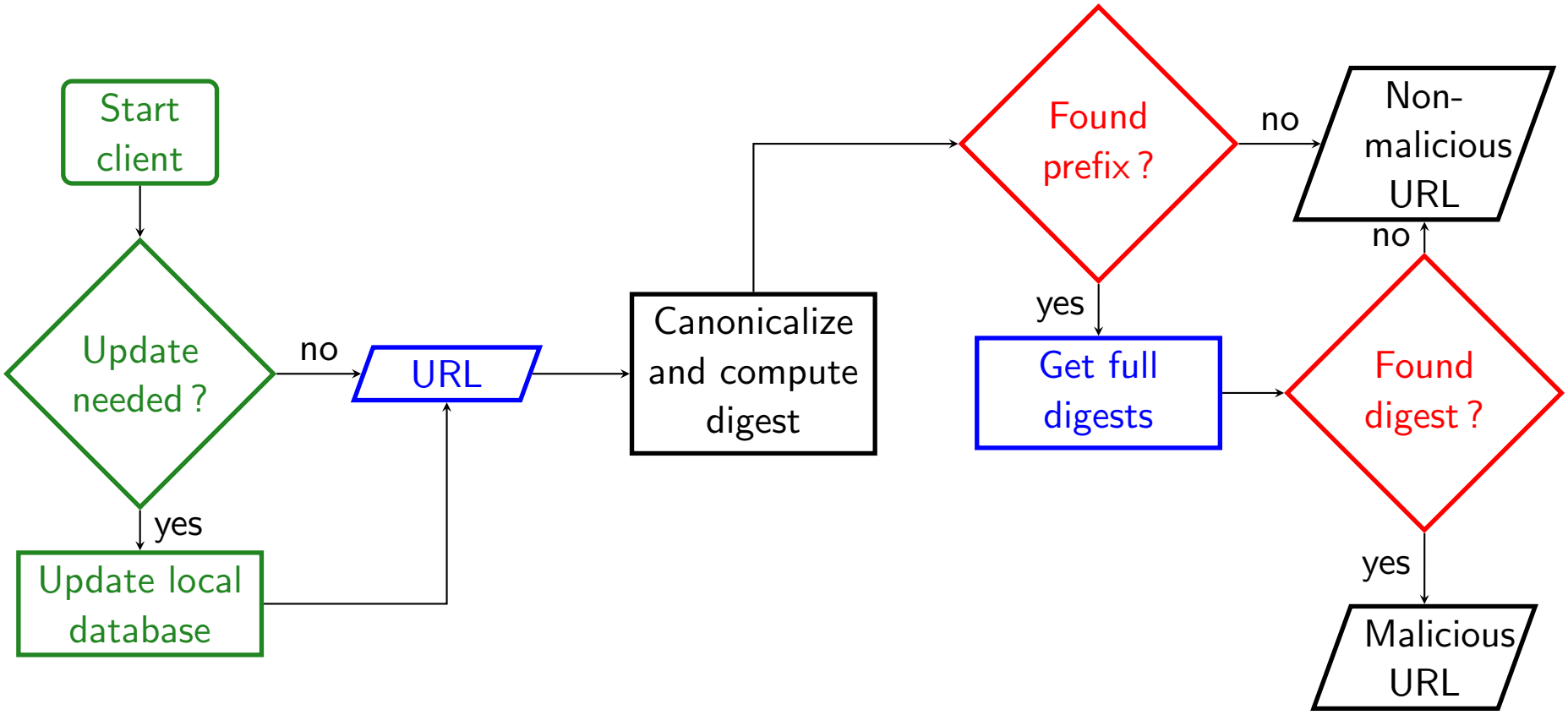
List name	**Description**	**#prefixes**
goog-malware-shavar	malware	317,807
goog-regtest-shavar	test file	29,667
goog-whitedomain-shavar	unused	1
googpub-phish-shavar	phishing	312,621

- **Environ $\approx$ 650000 entrées au total.**

- On ne travaille pas avec directement sur les URLs. On utilise les 4 premiers octets de l'empreinte SHA-256 (32 octets).

```
Prefix32(SHA256(www.example.com/))=0xd59cc9d3
```

Safe Browsing API v3



Why 32-bit prefixes ?

Code optimisation

		Structure (Mo)			
		Codage Delta		Filtre de Bloom	
Préfixe (bits)	**Don. brutes**	**taille**	**Compr.**	**taille**	**Compr.**
32	2. 5	1.3	1.9		0.8
64	5.1	3.9	1.3		1.7
80	6.4	5.1	1.2	3	2.1
128	10.2	8.9	1.1		3.4
256	20.3	19.1	1		6.7

Why 32-bit prefixes ?

Privacy

Year	# unique URLs (Google)	# of domains
2008	1 Trillion	177 Million
2012	30 Trillion	252 Million
2013	60 Trillion	271 Million

	M for URLs			M for domain		
ℓ (bits)	2008	2012	2013	2008	2012	2013
16	2^{28}	2^{28}	2^{29}	253	363	388
32	443	7541	14757	2	3	3
64	2	2	2	1	1	1
96	1	1	1	1	1	1

Yandex Safe Browsing API v3

List name	Description	#prefixes
goog-malware-shavar	malware	283,211
goog-mobile-only-malware-shavar	mobile malware	2,107
goog-phish-shavar	phishing	31,593
ydx-adult-shavar	adult website	434
ydx-adult-testing-shavar	test file	535
ydx-imgs-shavar	malicious image	0
ydx-malware-shavar	malware	283,211
ydx-mitb-masks-shavar	man-in-the-browser	87
ydx-mobile-only-malware-shavar	malware	2,107
ydx-phish-shavar	phishing	31,593
ydx-porno-hosts-top-shavar	pornography	99,990
ydx-sms-fraud-shavar	sms fraud	10,609
ydx-test-shavar	test file	0
ydx-yellow-shavar	shocking content	209
ydx-yellow-testing-shavar	test file	370
ydx-badcrxids-digestvar	.crx file ids	*
ydx-badbin-digestvar	malicious binary	*
ydx-mitb-uids	man-in-the-browser	*
ydx-badcrxids-testing-digestvar	test file	*

Re-identififcation

URL	32-bit prefix
https://cappris.inria.fr/project/events/	0x2929f0b1
https://cappris.inria.fr/project/	0xc99584e3
https://cappris.inria.fr/	0x192af851

- **Problème des valeurs positives :**
  - 1 match : 0x2929f0b1 $\rightarrow$ aucun problème.
  - 2 matches : 0xc99584e3 et 0x192af851 $\rightarrow$ re-identification.

- **Corrélations temporelles**

URL	32-bit prefix
https://cappris.inria.fr/	0x192af851
https://cappris.inria.fr/project/events/	0x2929f0b1

Caractéristiques des fichiers

- Reconstruction à partir de sources publiques.

		Malware list		Phishing list	
	list name	#matches	%matches	#matches	%matches
Google	goog-malware-shavar	18785	5.9	351	0.1
	googpub-phish-shavar	632	0.2	11155	3.5
Yandex	ydx-malware-shavar	44232	15.6	417	0.1
	ydx-adult-shavar	29	6.6	1	0.2
	ydx-mobile-only-malware-shavar	19	0.9	0	0
	ydx-phish-shavar	58	0.1	1568	4.9
	ydx-mitb-masks-shavar	20	22.9	0	0
	ydx-porno-hosts-top-shavar	1682	1.6	220	0.2
	ydx-sms-fraud-shavar	66	0.6	1	0.01
	ydx-yellow-shavar	43	20	1	0.4

Orphelins et faux-positifs

		#full hash per prefix				#Coll. with TopAlexa			
	list name	0	1	2	Total	0	1	2	Total
Google	goog-malware-shavar	36	317,759	12	317,807	0	572	0	572
	googpub-phish-shavar	123	312,494	4	312,621	0	88	0	88
Yandex	ydx-malware-shavar	4,184	279,015	12	283,211	73	2,614	0	2,687
	ydx-adult-shavar	184	250	0	434	38	43	0	81
	ydx-mobile-only-malware-shavar	130	1,977	0	2,107	2	22	0	24
	ydx-phish-shavar	31,325	268	0	31,593	22	0	0	22
	ydx-mitb-masks-shavar	87	0	0	87	2	0	0	2
	ydx-porno-hosts-top-shavar	240	99,750	0	99,990	43	17,541	0	17,584
	ydx-sms-fraud-shavar	10,162	447	0	10,609	76	3	0	79
	ydx-yellow-shavar	209	0	0	209	15	0	0	15

URLs remarquables

URL	matching decomposition	prefix
http ://fr.xhamster.com/user/video	fr.xhamster.com/	0xe4fdd86c
	xhamster.com/	0x3074e021
http ://nl.xhamster.com/user/video	nl.xhamster.com/	0xa95055ff
	xhamster.com/	0x3074e021
http ://m.mofos.com/user/login	m.mofos.com/	0x6e961650
	mofos.com/	0x00354501
http ://fr.xhamster.com/user/kmille	fr.xhamster.com/	0xe4fdd86c
	xhamster.com/	0x3074e021
http ://de.xhamster.com/user/video	de.xhamster.com/	0x0215bac9
	xhamster.com/	0x3074e021
http ://nl.xhamster.com/user/ppbbg	nl.xhamster.com/	0xa95055ff
	xhamster.com/	0x3074e021
http ://mobile.teenslovehugecocks.com/user/join	mobile.teenslovehugecocks.com/	0x585667a5
	teenslovehugecocks.com/	0x92824b5c
http ://nl.xhamster.com/user/photo	nl.xhamster.com/	0xa95055ff
	xhamster.com/	0x3074e021
http ://m.mofos.com/user/logout	m.mofos.com/	0x6e961650
	mofos.com/	0x00354501

Censure avec Yandex.Browser

secours-islamique.org

Conclusion

- Des URLs peuvent être re-identifiées par Google et Yandex.

- Les listes sont choisies **arbitrairement** :
  - reverse impossible. . .
  - orphelins,
  - justification des valeurs.

  On a besoin d'*accountability*.

- Solution (par défaut) : trivial *Private Information Retrieval* !
  - accountable
  - private