Analyse des heuristiques prédominants pour la détection des sites de phishing

Sommaire

Le Chapitre 2 a introduit les attaques de phishing et leur principal vecteur de diffusion : le spam. Nous y avons notamment détaillé les principales caractéristiques d'un site de phishing ainsi que les moyens de détection/prévention associés. Dans notre volonté de nous focaliser sur le poste client de l'Internaute, nous avons vu qu'un moyen de détection facile d'accès qui leur est proposé s'installe/se configure dans le navigateur web. Il s'agit des barres d'outils anti-phishing qui se basent sur des listes noires (et éventuellement blanches) et/ou des tests heuristiques.

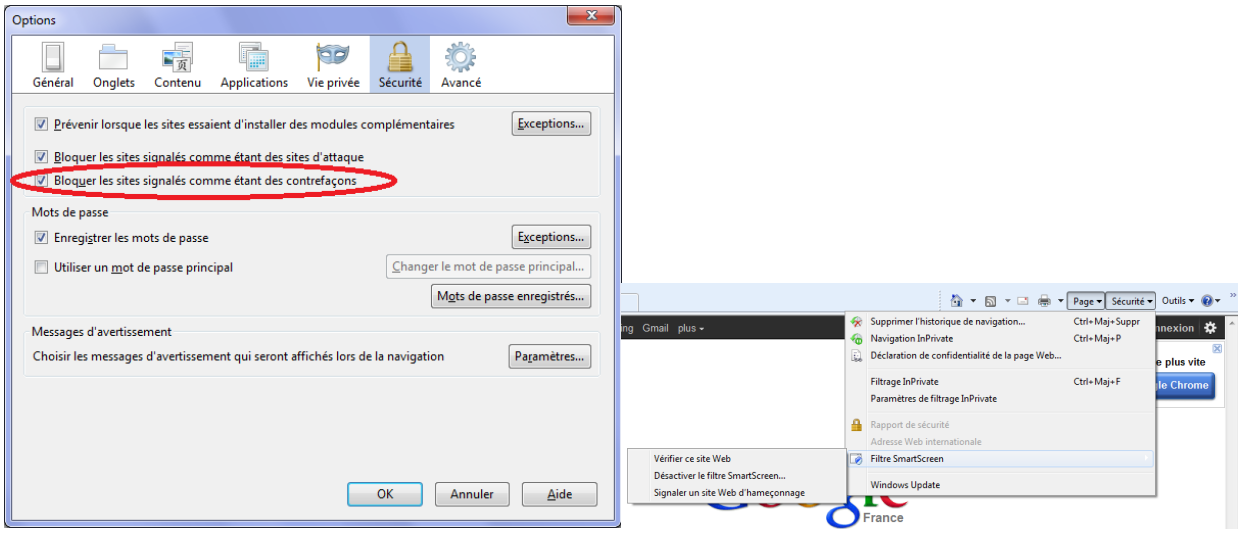
Dans ce chapitre, nous avons souhaité examiner plus avant ces barres anti-phishing. En particulier, nous nous sommes intéressés aux tests heuristiques qu'elles utilisent, afin d'en évaluer l'efficacité/la pérennité à distinguer les sites légitimes des sites contrefaits. Pour ce faire, nous nous sommes appuyés sur la conception de notre propre barre d'outils anti-phishing - nommée *Phishark* -, exclusivement basée sur les tests heuristiques.

En section 3.1, nous démarrons par une introduction qui place le contexte de notre étude et la situe par rapport aux travaux similaires existants. La section 3.2 explique ensuite la manière dont notre moteur de détection prend sa décision de légitimité/contrefaçon. Puis, la section 3.3 détaille les conditions d'expérimentation des différentes phases de tests. Dans la continuité - basés sur notre analyse des caractéristiques des sites de phishing vus dans le Chapitre 2 - nous détaillons en section 3.4 les 20 tests heuristiques étudiés ainsi que les seuils de classification légitime/contrefait associés. Nous y abordons également la barre d'outils anti-phishing développée pour évaluer l'efficacité des heuristiques. La section 3.5 s'attache ensuite à vérifier l'efficacité des heuristiques et seuils de décision choisis. En marge, nous comparons les performances de la barre développée aux barres d'outils anti-phishing les plus courantes. Puis, en section 3.6 nous déduisons de ces tests les heuristiques les plus pertinents pour l'identification des sites légitimes et contrefaits. Nous discutons également de leur pérennité en section 3.7. Enfin, nous terminons par un examen des problèmes rencontrés en section 3.8. Les tests effectués dans ce chapitre portent sur 650 URLs légitimes et 950 URLs de phishing.

Ce chapitre fait partie de nos contributions : les résultats de notre étude ont été publiés et présentés à la conférence *Sécurité des Architectures Réseaux - Sécurité des Systèmes d'information* (SAR-SSI) en Mai 2011 [GGL11].

3.1 Introduction à l'approche développée

Parmi les moyens de détection/protection du phishing, il y en a un qui est particulièrement facile d'accès aux Internautes : la barre d'outils anti-phishing. A l'origine, celle-ci se présentait sous forme d'un plug-in additionnel à installer au sein du navigateur. Au cours des dernières années, devant la prolifération des sites de phishing, les éditeurs ont intégré ces fonctionnalités nativement au sein de leurs navigateurs (p.ex. depuis la version 3 pour Mozilla Firefox, ou depuis la version 7 pour Microsoft Internet Explorer - cf. figure 3.1). Néanmoins, d'autres barres alternatives que celles proposées par les navigateurs - plus ou moins maintenues - peuvent être utilisées (p.ex. Spoofguard [CLT+04], Spoofstick, Netcraft, WOT (pour *Web of Trust*), etc. - cf. aperçus visuels en figure 2.14 de la section 2.4.3.4). Les logiciels anti-virus les plus courants (p.ex. BitDefender, Symantec Norton, McAfee Site Advisor, etc.) proposent également nativement des barres anti-phishing qui s'intègrent automatiquement (c.-à-d. au moment de l'installation de l'anti-virus) dans le navigateur. Précisons enfin que des attaquants explorent également ce filon, en diffusant régulièrement de fausses barres anti-phishing sur le web.



(a) Firefox 3.5

(b) Internet Explorer 8

FIGURE 3.1 – Configuration des fonctionnalités anti-phishing dans deux navigateurs webs courants : Mozilla Firefox et Microsoft Internet Explorer

Les barres d'outils anti-phishing basent leur détection sur l'utilisation de listes noires et/ou tests heuristiques. Précisons que ces deux techniques sont toutes autant utilisées. Elles ont d'ailleurs donné lieu à de nombreux articles et, bien que les avis divergent, il apparaît difficile de trancher définitivement en faveur de l'une ou l'autre de ces techniques.

L'utilisation des listes noires s'avère relativement contradictoire avec l'une des caractéristiques principales des sites de phishing, à savoir : leur durée de vie très courte (cf. section 2.3). En effet, bien que les listes noires s'avèrent plus exactes dans leur détection (c.-à-d. à priori, elles ne présentent pas de faux-négatifs), elles n'en sont pas moins incomplètes. De plus, elles nécessitent une mise à jour en temps réel ou presque (c.-à-d. dès l'apparition du site de phishing). Enfin, que la liste noire soit stockée côté client et/ou récupérée/testée depuis Internet, elle est une cible providentielle pour les attaquants (p.ex. via une attaque de type *Man-in-the-Middle*).

A contrario, les tests heuristiques semblent moins vulnérables car ils fonctionnent de manière autonome. De plus, ils ne nécessitent pas de mises à jour fréquentes. Néanmoins, croire que définir des tests heuristiques à un instant t peut s'avérer suffisant est un leurre. En effet, leur dimension plus "statique" leur confère un degré de péremption non négligeable. Dès lors que les attaquants ont connaissance des tests heuristiques et/ou seuils de détection associés, ils peuvent essayer d'adapter leurs contrefaçons de sites webs afin qu'elles leurrent la détection.

D'où la forte association des deux familles de techniques (listes noires et tests heuristiques) dans les barres anti-phishing. L'étude menée par Sheng et al. [SWW+09] a d'ailleurs démontré que leur utilisation combinée était un facteur indispensable à une détection efficace et optimisée.

Cette même étude montre notamment que 63% des 191 campagnes de phishing qu'ils ont étudiées étaient terminées au bout de 2 heures, alors que les listes noires testées (utilisées par McAfee Site Advisor, Symantec Norton, Netcraft, Internet Explorer, Chrome ou Firefox) n'étaient capables d'en détecter que 20% à l'instant *t=0*. Ces mêmes listes noires sont également mises à jour dans des délais très variables, aboutissant à une détection de 43 à 87% des URLs de phishing au bout de 12 heures. En complément, une étude menée par Kumaraguru et al. [KCA+09] a démontré qu'après réception d'un email de phishing, au moins 50% des victimes potentielles accédaient au site contrefait dans les 2 premières heures, et 90% dans les 8 heures.

Diverses suggestions/solutions sont proposées pour améliorer l'efficacité des listes noires. On peut notamment citer que Sheng et al. [SWW+09] proposent de générer celles-ci à partir des filtres anti-spam, partant du principe que les attaques de phishing sont majoritairement véhiculées par les emails. De leur côté, Prakash et al [PKKG10] ont développé un outil visant à prédire des listes noires d'URLs depuis une liste noire source, grâce à l'utilisation d'heuristiques (p.ex. changement de TLD, dérivation de nom de fichier, etc.). Leur théorie s'appuie sur le fait que les attaquants génèrent souvent de multiples URLs contrefaites très ressemblantes (p.ex. `http://g00gle.hdfree.in/1.html` et `http://g00gle.hdfree.in/2.html`, ou `http://e-baltikums-mailbox.com/different.files/geoga.html` et `http://e-baltikums.info/different.files/arta.html`).

Toutefois, il n'en demeure pas moins que les listes noires semblent majoritairement inefficaces au moment où la probabilité d'accéder au site contrefait est la plus forte. Nous avons donc choisi de nous intéresser aux heuristiques sur lesquels repose la détection dans cette phase clé.

Plus particulièrement, nous nous sommes intéressés à évaluer l'efficacité des tests heuristiques, qu'ils portent sur l'analyse de l'URL et/ou sur l'étude du contenu de la page web visitée. Pour ce faire, nous nous sommes appuyés sur le développement de notre propre barre anti-phishing - *Phishark* - conçue tel un plug-in pour Mozilla Firefox, premier navigateur web open source [web] au niveau mondial, et N° 1 en Europe devant Internet Explorer [Reu11].

Nous y avons intégré 20 tests heuristiques répartis en 6 catégories. Trois catégories s'intéressent à l'analyse de l'URL visitée, tandis que les trois autres se focalisent sur l'étude du contenu de la page web.

Le moteur de détection a été étalonné, puis vérifié, sur des jeux d'URLs légitimes/contrefaites différents. A partir des résultats obtenus lors de la vérification, nous avons déterminé les catégories de tests heuristiques les plus déterminantes (c.-à-d. celles conduisant à des scores négatifs pour les listes noires, et celles conduisant à un score positif pour les listes blanches) pour la différenciation des sites légitimes et contrefaits.

Positionnement par rapport aux travaux similaires : Les travaux précédents qui s'apparentent le plus à notre étude ont été amorcés dans le Chapitre 2. Nous les détaillons plus avant ci-après afin de nous positionner :

- L'étude de Zhang et al. [ZECH07] s'est intéressée à évaluer/comparer l'efficacité de 10 barres d'outils anti-phishing, dont la décision de légitimité s'articule principalement autour de blacklists. Sur les échantillons d'URLs testées, un manque de réactivité dans la mise à jour des blacklists utilisées est apparu. Sur ces points, nos études/avis se rejoignent. Néanmoins, leur étude n'apporte aucune mesure d'efficacité des tests heuristiques (les uns par rapport aux autres, ou vs. les blacklists).

- Les travaux de Ma et al. [MSSV09] se sont focalisés sur une étude de l'URL pour en déduire une méthode de classification des sites légitimes/contrefaits. Plusieurs des critères pertinents qu'ils utilisent ont été retrouvés dans une partie amont de notre étude (cf. section 3.4.1.2). Cette dernière nous a permis d'aboutir à une liste de triplets déterminants que nous utilisons lors de notre analyse de l'URL visitée. A contrario de notre étude, l'utilisation de ces critères est le socle de leur détection/classification. Un autre point de divergence de nos travaux réside dans le fait qu'ils ne s'intéressent pas à l'analyse du code source HTML.

- L'étude menée par Ludl et al [LMKK07] est probablement celle qui s'apparente le plus à nos travaux. Elle s'est intéressée à évaluer l'efficacité de deux blacklists - utilisées par Mozilla Firefox et Internet Explorer - pour la détection des sites de phishing, ainsi qu'à distinguer les caractéristiques prédominantes pouvant amener à une décision de légitimité/contrefaçon d'un site. Leurs travaux aboutissent à l'élaboration d'un arbre de décision qui s'articule aussi bien autour de l'analyse de l'URL que du contenu de la page web. Néanmoins il apparaît difficile de vraiment comparer nos deux études puisque leurs tests heuristiques sont peu détaillés (en terme de contenus étudiés et/ou scores décisionnels associés). On peut néanmoins dire que pour les sites légitimes, leurs résultats portent sur davantage d'URLs que nous en avons testés. Leur taux de faux-positifs obtenu est de 0.43%. Sur les pages de phishing, nos bases d'URLs sont de tailles comparables, mais leur taux de faux-négatifs est relativement important : 16.97%. Il semble également que nos tests heuristiques qui s'intéressent à l'URL sont plus nombreux. Enfin, leur étude ne discute pas de l'éventuelle pérennité des tests utilisés.

3.2 Décision de légitimité/contrefaçon

La décision de légitimité/contrefaçon prise par le moteur de détection de notre barre d'outils baptisée *Phishark*, repose sur le calcul d'un score global établi en additionnant - avec un poids équivalent - l'ensemble des scores obtenus pour chaque heuristique :

$$\text{Score global (p)} = \sum_{i=1}^{20} \text{Score(i)}$$

où *(p)* représente la page analysée/visitée dans le navigateur
et *(i)* le numéro d'heuristique

Si le moteur de détection ne peut aboutir à une décision (c.-à-d. le score global est = 0), le site visité est considéré comme *Risqué* (notre barre d'outils affiche *Risky Site*). Si le score global est > 0, le site visité est indiqué *Légitime* (notre barre d'outils affiche *Legitimate Site*). Sinon (c.-à-d. si le score global est < 0), le site visité est décidé *Contrefait* (notre barre d'outils affiche *Phishing Site*).

Les seuils de décision définitifs associés à chaque heuristique sont déterminés au travers de la combinaison de 3 facteurs : 1/ d'éventuels seuils de décision relevés au travers de précédents travaux qui ont servi de base à notre étalonnage, 2/ de nos études complémentaires approfondies sur les pages légitimes/contrefaites, qui ont abouti à des critères supplémentaires et/ou des seuils affinés, 3/ ou encore de notre phase d'étalonnage du moteur de détection qui a permis d'arriver au meilleur compromis (c.-à-d. à un minimum de décisions erronées) dans la détection des sites légitimes/contrefaits.

3.3 Conditions d'expérimentation

Afin d'évaluer la pertinence des tests heuristiques utilisés par les barres d'outils anti-phishing, nous avons réalisé deux types de tests : 1/ des tests portant sur des sites légitimes, et 2/ des tests portant sur des sites de phishing.

Ces deux types de tests ont été réalisés par 3 phases différentes : 1/ une première phase dite d'étalonnage qui a permis de définir des seuils de détection optimisés, 2/ une deuxième phase de vérification des performances du moteur de détection *Phishark*, en comparaison avec d'autres barres d'outils anti-phishing courantes, et 3/ une troisième phase d'identification des heuristiques déterminants[1].

Précisons que pour améliorer la lisibilité des explications données dans ce chapitre, les listes d'URLs légitimes sont ultérieurement nommées *whitelist*, tandis que les listes d'URLs contrefaites (c.-à-d. de phishing) sont désormais nommées *blacklist*.

3.3.1 Établissement des listes d'URLs

Une des difficultés premières de notre étude réside dans la possibilité de tester des sites de phishing. En effet, de par leur durée de vie très courte, les sites de phishing doivent être testés dès leur apparition. Pour sélectionner ces URLs, nous nous sommes appuyés sur les sites de l'APWG [apw] et de Phishtank [phi] qui délivrent des bases d'URLs de sites contrefaits. Nous avons choisi de sélectionner des sites de phishing "validés" à 100%. En effet, les bases de données communiquées par l'APWG et Phishtank indiquent un niveau de confiance (c.-à-d. il est avéré et vérifié que l'URL présentée est un site de phishing), variant de 50 à 100% pour les URLs mises à disposition. Nous avons pris soin de sélectionner - pour chaque phase de tests - des URLs de phishing d'aspects très variables, répondant à l'ensemble des caractéristiques exposées en section 2.2. Précisons toutefois que les URLs de phishing de type HTTPS sont rares.

De plus, ces URLs de phishing ont été récupérées et aussitôt testées par blocs de 50 à 100 URLs maximum, pour limiter la quantité d'URLs périmées. Même en procédant ainsi, nous avons eu des taux de pertes atteignant jusqu'à 7% par jeu d'URLs récupérées.

Les URLs de phishing collectées - récupérées entre Janvier et Septembre 2010 - usurpent majoritairement des sites bancaires (p.ex. Bank of America, Paypal, Chase, etc.), des sites d'e-commerce (p.ex. eBay, etc.), des sites d'email (p.ex. Hotmail), des sites de réseaux sociaux (p.ex. Facebook) ou autres (p.ex. jeux en ligne avec RuneScape).

Les whitelists ont quant à elles été générées à partir de diverses sources. On peut notamment citer le Top 1000 des sites les plus visités délivré par Google [Gooc], le Top 500 des sites les plus visités publié par Alexa [Al], le Top 100 des sites les plus populaires édité par Netcraft [Net], la whitelist proposée par Google [Goob], des URLs de sites bancaires récupérés grâce au site Levoyageur [lev], etc.

Nous avons pris soin - pour chaque phase de tests - de multiplier les secteurs d'activité, le type de site visité (HTTP et HTTPS), le langage des pages webs, les types d'URLs sélectionnées (FQDN simple ou suivi de multiples niveaux d'arborescence), ou encore les TLDs (Top-Level Domain).

L'ensemble des URLs légitimes utilisées pour cette étude ont été collectées entre Avril 2009 et Septembre 2010.

Bien que les sites de phishing visent à usurper en grande majorité des sites de login, notre sélection d'URLs légitimes ne s'est pas restreint à cette catégorie de sites. En effet, notre moteur de détection s'exécute à chaque URL visitée. Il se doit donc de délivrer les bonnes décisions quel que soit le type de site web visité, d'où la diversité d'URLs sélectionnées.

3.3.2 Phase d'étalonnage

Une des phases les plus délicates de l'étude développée dans ce chapitre réside dans l'ajustement des seuils de décision associés à chaque heuristique.

Pour ce faire, la phase d'étalonnage s'est déroulée en deux temps : 1/ nous avons cherché à définir les seuils de décision optimum pour les sites de phishing testés (c.-à-d. un maximum de décisions

1. Notons que cette troisième phase a pour vocation d'aider à affiner les heuristiques et seuils de détection associés, lors d'une étude ultérieure (cf. Chapitre 7).

négatives), puis 2/ nous avons cherché à définir les seuils de décision optimum pour les sites légitimes testés (c.-à-d. un maximum de décisions positives).

A partir de ces résultats obtenus en 2 temps, nous avons ensuite réalisé un compromis pour aboutir à des seuils de détection optimum pour les deux types d'URLs : légitimes et contrefaites. En effet, chaque modification de seuil peut affecter une liste positivement et l'autre négativement. Il a donc fallu procéder à plusieurs séries de tests successifs avant d'aboutir aux tableaux optimum présentés pour chaque liste (whitelist et blacklist) en section 3.4.

Cette phase d'étalonnage a porté sur une whitelist de 150 URLs, et une blacklist de 200 URLs issues du site de Phishtank.

3.3.3 Phase de vérification et de comparaison aux autres barres d'outils

Dans cette deuxième phase, nous avons voulu vérifier l'efficacité de notre moteur de détection. En marge, nous en avons profité pour le confronter aux performances des barres d'outils anti-phishing les plus courantes.

Les tests réalisés dans cette phase ont porté sur une whitelist de 500 nouvelles URLs, et une blacklist de 520 nouvelles URLs issues des sites de Phishtank (à hauteur de 69.81%) et de l'APWG (à hauteur de 30.19%).

Barres anti-phishing sélectionnées : A partir de travaux précédents [ZHC07] [WMG06] et d'une étude que nous avons menée sur les barres anti-phishing les plus courantes, nous avons sélectionné 4 barres d'outils qui apparaissent comme les mieux classées en terme de performance (c.-à-d. qui délivrent les meilleurs taux de détection avec un minimum de FPR/FNR) et les mieux maintenues. Ainsi, nous avons comparé les performances de *Phishark* avec :

- Le navigateur web ***Mozilla Firefox v.3.6.7*** qui intègre un moteur de détection anti-phishing basé sur des listes blanches/noires (parmi lesquelles celles de Google et de Phishtank), la liste noire étant rafraîchie toutes les 30 minutes [Moz]. D'après leurs travaux, Zhang et al. [ZECH07] présument également que le moteur de détection de Mozilla Firefox utilise des tests heuristiques.
- La barre anti-phishing ***Netcraft v.1.4.1.5*** qui s'installe telle un plug-in dans le navigateur web Mozilla Firefox. Elle appuie sa détection sur l'utilisation des listes blanches/noires propres à Netcraft (auxquelles l'Internaute peut contribuer) : à chaque page web visitée, la barre d'outils envoie l'URL de la page à contrôler aux serveurs Netcraft pour vérification. Cet envoi est non sécurisé [AN08]. En complément, la barre anti-phishing utilise des tests heuristiques tels que la vérification de l'adresse IP du site, le pays hébergeur, le nom du serveur web, l'âge du site, etc.
- Le navigateur web ***Internet Explorer v.8.0.601.18702*** qui intègre un moteur de détection anti-phishing nommé *SmartScreen Filter*. Celui-ci utilise des listes blanches/noires maintenues par Microsoft. La liste noire provient du service en ligne *Microsoft URL Reputation Service* qui s'appuie sur des listes noires externes à Microsoft, auxquelles les Internautes contribuent. En complément, elle utilise des tests heuristiques [Mic].
- La barre ***Web of Trust (WOT) v.20100503***, qui s'installe telle une extension dans le navigateur web (au choix : Mozilla Firefox, Internet Explorer, Safari, Opera ou Google Chrome). Elle utilise des tests heuristiques (p.ex. la popularité du site web, la localisation du serveur web et sa réputation, etc.), ainsi qu'une liste noire qui inclut des URLs délivrées par Phishtank [Ser06].

3.3.4 Phase d'identification des heuristiques déterminants

Dans cette troisième phase, nous avons cherché à identifier les heuristiques prédominants qui permettent de distinguer un site légitime d'un site contrefait, et réciproquement.

Pour ce faire, nous avons regardé les scores attribués à chaque catégorie d'heuristiques (les catégories sont détaillées en section 3.4), pour l'ensemble des URLs testées. Le score de chaque catégorie d'heuristiques (H) - exprimé en pourcentage - est ainsi déterminé en calculant la proportion d'URLs qui présentent un score négatif (pour les blacklists (BL)) ou positif (pour les whitelists (WL)), parmi l'ensemble des URLs testées :

Tableau 3.1 – Heuristiques implémentés

Catégorie	Nº	Heuristique
Points et caractères spéciaux	1	Nombre de points (.) dans l'URL
	2	Nombre d'arobas (@) dans l'URL
	3	Nombre de double slash (//) dans l'URL
	4	Présence d'une adresse IP dans le FQDN
	5	Présence d'un numéro de port dans le FQDN
Triplets et mots-clés (dits de phishing)	6	Nombre de triplets dans le FQDN
	7	Nombre de triplets dans l'arborescence de l'URL (hors FQDN)
	8	Nombre de mots-clés dans l'arborescence de l'URL (hors FQDN)
TLD	9	Présence d'un TLD "sensible" dans le FQDN
	10	Présence d'un TLD "sensible" dans l'arborescence de l'URL (hors FQDN)
	11	Comparaison entre TLD et pays hébergeur du site
Code source HTML	12	Evaluation de la balise de Titre
	13	Evaluation de balise Formulaire
	14	Evaluation de lien Image
	15	Evaluation d'autre lien
Page de Login	16	HTTPS et zones de login
Autres balises HTML	17	Evaluation de balise de Description Meta
	18	Evaluation de balise de mots-clés Meta
	19	Evaluation de balise Script
	20	Evaluation de balise Link

$$\text{Score } H(x)_{BL} = \frac{\sum_{i=1}^{n}\left(H(i)<0\right)}{n}\cdot 100 \quad \text{et} \quad \text{Score } H(x)_{WL} = \frac{\sum_{i=1}^{n}\left(H(i)>0\right)}{n}\cdot 100$$

où *(x)* représente la catégorie d'heuristique étudiée
et *(n)* le nombre d'URLs testées

Les tests réalisés ici ont porté sur une whitelist de 230 URLs (extraites des 500 URLs utilisées lors de la phase de vérification), et sur une blacklist de 230 nouvelles URLs.

3.3.5 Environnement

Pour *Phishark*, les 3 phases ont été réalisées depuis une seule et même machine ayant pour caractéristiques : double processeur Intel® Core™2, 0,99 Go. de RAM, système d'exploitation Microsoft Windows XP Professional, version 2002, Service Pack 3, et navigateur web Mozilla Firefox v.3.6.7.

Pour la phase de comparaison des performances, afin de garantir les mêmes conditions d'expérimentation à toutes les barres anti-phishing, chaque barre sélectionnée a été installée sur une machine de mêmes caractéristiques (exceptée une machine où le navigateur web est différent, afin de tester la fonctionnalité anti-phishing d'*Internet Explorer*).

Ainsi la phase de comparaison a été réalisée en simultané pour chaque barre d'outils (c.-à-d. 1 URL testée au même moment par les 5 barres).

3.4 Phase d'étalonnage : Heuristiques étudiés et seuils de détection

Basés sur l'analyse des caractéristiques des pages de phishing vue en section 2.2, et des travaux précédents [ZECH07] [ZHC07] [HYM09], nous avons défini et implémenté 20 tests heuristiques qui se décomposent en 6 catégories (cf. tableau 3.1) :

- **Points et caractères spéciaux**
- **Triplets et mots-clés (dits de phishing)**
- **TLD**
- **Code source HTML**
- **Page de login**
- **Autres balises HTML**

Ces heuristiques ont été choisis à partir :
- D'études précédentes qui ont fait apparaître les principales caractéristiques à analyser et/ou seuils de décision associés. Ces travaux - qui portent en grande majorité sur l'URL - sont cités au fur et à mesure des détails donnés ci-après à propos de chaque heuristique.
- Et d'études complémentaires approfondies que nous avons menées sur l'analyse des caractéristiques des pages légitimes/contrefaites, tant au niveau de l'URL que du contenu du code source des pages HTML.

Précisons que quel que soit le seuil de décision retenu pour chaque heuristique, il existe toujours de nombreux cas de faux-positifs et/ou faux-négatifs résiduels associés. Seul l'ensemble des heuristiques cumulés - pour arriver au score global de décision - peut limiter ces cas. En effet, plus il y a de scores heuristiques individuels négatifs, plus la probabilité d'être en présence d'un site contrefait est forte. Et inversement.

3.4.1 URL

Les sites de phishing reposent en partie sur l'utilisation d'une URL alternative. Une piste de détection évidente est donc l'analyse de cette URL.

Dans notre étude, les heuristiques qui s'y rapportent sont regroupés dans les 3 catégories suivantes : *Points et caractères spéciaux*, *Triplets et mots-clés (dits de phishing)*, et *TLD* (cf. tableau 3.1).

3.4.1.1 Catégorie Points et caractères spéciaux

3.4.1.1.1 Heuristique N° 1 : Nombre de points (.) dans l'URL L'étude de Zhang et al. [ZHC07] indique que les URLs de phishing sont généralement constituées d'un minimum de 5 points. En effet, pour leurrer les utilisateurs, les attaquants ont tendance a faire référence au FQDN légitime dans l'arborescence de l'URL. Par exemple, l'URL `http://www.neural-net.ca/store/images/webscr/1/www.paypal.co.uk/` qui usurpe le site Paypal et a pour FQDN `www.neural-net.ca`, rappelle le FQDN légitime `www.paypal.co.uk` dans son arborescence, ce qui introduit des points supplémentaires (vs. une URL "classique").

Une URL légitime contient quant à elle généralement un minimum de 2 ou 3 points (p.ex. `www.mondomaine.com` ou `www.mondomaine.com/dossier1/dossier2/fichier.html`). Il est en effet plus rare de voir des URLs de phishing uniquement constituées d'un FQDN.

Par conséquent, nous avons choisi la valeur de 3 points comme valeur charnière de la décision de légitimité d'un site (cf. tableau 3.3). Jusqu'à cette valeur, le score de l'heuristique est positif ou nul. Au-delà, il est négatif. Précisons que cet heuristique s'applique à l'intégralité de l'URL.

La phase d'étalonnage démontre que les seuils optimum sont identiques sur blacklist et whitelist (cf. tableau 3.4).

Des exemples de faux-positifs / faux-négatifs résiduels :
- L'URL légitime du site de login Hotmail `http://login.live.com/login.srf?wa=wsignin1.0&rpsnv=11&ct=1281707336&rver=6.0.5285.0&wp=MBI&wreply=http:%2F%2Fmail.live.com%2Fdefault.aspx&lc=2057&id=64855&mkt=en-gb` comporte 10 points.
- L'URL légitime du site de login SFR `https://www.sfr.fr/cas/login?service=https%3A%2F%2Fsfr-messagerie.services.sfr.fr%2Fwebmail%2Fj_spring_cas_security_check` comprend 5 points.
- L'URL contrefaite `http://gomezcomunidade.t35.com/` ne contient que 2 points.

3.4.1.1.2 Heuristiques N° 2 et 3 : Nombre d'arobas (@) et de double slash (//) dans l'URL Les caractères spéciaux tels que l'arobas (@) - situé dans le FQDN -, ou le double slash (//) - situé dans l'arborescence de l'URL - sont parfois utilisés par les attaquants, soit pour procéder à d'éventuelles redirections, soit pour leurrer l'utilisateur en faisant apparaître le FQDN légitime (cf. section 2.2).

Par conséquent, dès lors qu'un caractère de ce type est détecté lors de l'analyse, le score de l'heuristique correspondant est négatif. Sinon, il est positif. Ces deux heuristiques s'appliquent à l'intégralité de l'URL, au-delà de l'indication de protocole - c.-à-d. au delà du `http://` - pour la recherche de double

slash (cf. tableau 3.3).

La phase d'étalonnage sur blacklist et whitelist démontre que les seuils optimum sont identiques pour la présence d'arobas .

A contrario, il arrive de trouver le double slash à de multiples reprises dans les URLs de phishing (cf. tableau 3.4), souvent afin de rappeler le FQDN légitime.

On trouve également des double slash dans les URLs légitimes, typiquement pour des redirections de pages de login (parfois ces redirections sont masquées via des techniques d'encodage, p.ex. `https://` peut être remplacé par `https%3A%2F%2F` avec un encodage ASCII). Toutefois, la majorité des URLs légitimes qui font appel à ces techniques de redirection renvoient généralement vers un FQDN appartenant au même domaine. Nous avons donc pris cette notion en compte pour déterminer le seuil de décision final (cf. tableau 3.5).

Des exemples d'URLs faisant appel à ces techniques de redirection :
- L'URL légitime du site de login Orange `http://id.orange.fr/auth_user/bin/auth0user.cgi?date=1276854069&skey=0fe469df5d9227a79b29390183013939&service=communiquer&url=http://webmail1j.orange.fr/webmail/fr_FR/welcome_freeUrl.html` utilise le (//) pour rediriger vers le FQDN `webmail1j.orange.fr` appartenant au même domaine.
- L'URL légitime du site de login Wachovia `https://www.wachovia.com/stateselector?referring_page=https://www.wachovia.com/foundation/v/index.jsp?vgnextoid=06ce9e05d1674210VgnVCM200000627d6fa2RCRD&product_code=CHK` utilise le (//) pour rediriger vers le même FQDN.
- L'URL contrefaite `http://desoskiz.org/personalbankingalertnotification@hsbc.co.uk/index.php` utilise l'(@) pour passer des attributs et faire apparaître le FQDN légitime.
- L'URL contrefaite `http://www.amazon.com:fvthsgbljhfcs83infoupdate@69.10.142.34` [CG06] semblerait indiquer pour un utilisateur non-averti qu'il est dirigé vers le FQDN `www.amazon.com`. Néanmoins, l'utilisation du (@) le redirige vers le serveur web de l'attaquant, qui a pour adresse IP : `69.10.142.34`.

3.4.1.1.3 Heuristique N° 4 : Présence d'une adresse IP dans le FQDN Pour éviter qu'un FQDN différent du FQDN légitime n'attire l'attention de l'Internaute, les attaquants préfèrent parfois le remplacer par une adresse IP [CG06]. Nous constatons que ce remplacement est d'ailleurs souvent couplé à un rappel du domaine ou FQDN légitime dans l'arborescence de l'URL, afin de maximiser les chances de leurrer l'utilisateur.

A contrario, ce cas de figure se présente très rarement sur URLs légitimes (une adresse IP est parfois utilisée dans les URLs qui utilisent le protocole FTP, mais celles-ci sont rarement la cible d'attaques de phishing).

Cet heuristique s'applique à l'analyse du FQDN exclusivement. A l'examen de l'URL, dès lors qu'une adresse IP est détectée en lieu et place d'un FQDN (cf. tableau 3.3), nous attribuons un score négatif à l'heuristique correspondant. Dans le cas contraire, le score assigné est nul. Précisons toutefois que rien n'empêche un utilisateur d'accéder à un site légitime via son adresse IP. Le score négatif attribué ici se doit donc d'être limité pour pallier ce cas de figure.

La phase d'étalonnage démontre que les seuils optimum de ce critère sont identiques sur blacklist et whitelist (cf. tableau 3.4).

Des exemples d'URLs contrefaites qui utilisent une adresse IP en lieu et place d'un FQDN :
- L'URL contrefaite `http://216.139.111.51/~hsbc/hsbc/`, qui usurpe le site HSBC, prend soin de rappeler le domaine légitime dans l'arborescence de l'URL.
- L'URL contrefaite `http://221.5.225.151/www.paypal.com.it.cgi-bin.webscr.cmd.login.run/index.htm`, qui usurpe le site Paypal, prend soin de rappeler le FQDN légitime dans l'arborescence de l'URL.

3.4.1.1.4 Heuristique N° 5 : Présence d'un numéro de port dans le FQDN L'APWG [APW10] indique que les URLs de phishing utilisent parfois des ports alternatifs - en association avec le FQDN - à ceux traditionnellement utilisés par les protocoles (exemples de ports standards : 80 pour le HTTP, 443 pour

Tableau 3.2 – Liste des 240 triplets

".at"	".au"	".br"	".ch"	".cn"	".co"	".de"	".eb"	".ed"	".es"	".eu"	".go"	".il"	".in"	".iv"	".mo"	".ms"	".ne"	".nl"	".nz"
".or"	".rr"	".ru"	".tk"	".to"	".us"	".ya"	"0.n"	"10."	"1nc"	"a.g"	"a.o"	"a.u"	"adf"	"adu"	"aeo"	"aho"	"ail"	"an."	"ank"
"arc"	"art"	"asi"	"at."	"au."	"aud"	"aue"	"b.c"	"ban"	"bay"	"bes"	"bmw"	"br."	"c.u"	"cas"	"ch."	"cit"	"cn."	"co."	"com"
"cor"	"cou"	"cro"	"cs."	"d.o"	"d.u"	"dco"	"dcr"	"de."	"deb"	"dru"	"du."	"e.i"	"e.o"	"e.u"	"eat"	"ear"	"eba"	"ebt"	"eca"
"ech"	"eco"	"edu"	"ej."	"eo."	"eof"	"er."	"erc"	"ers"	"es."	"et."	"eu."	"ews"	"exf"	"ez."	"f.c"	"fan"	"fil"	"fo."	"gen"
"gir"	"gov"	"h.u"	"hoo"	"iae"	"iau"	"ics"	"ieo"	"if."	"ij."	"ik."	"il."	"ilm"	"inc"	"inf"	"ino"	"int"	"iq."	"irl"	"ity"
"iw."	"j.c"	"jou"	"l.c"	"l.o"	"l.u"	"lan"	"lms"	"loa"	"m.b"	"m.c"	"m.e"	"mai"	"mer"	"mon"	"msn"	"n.c"	"n.u"	"nal"	"nc."
"nc0"	"nco"	"net"	"new"	"nfo"	"nk."	"nl."	"nlo"	"no."	"ns."	"nst"	"nte"	"nty"	"nux"	"nz."	"o.c"	"o.n"	"o.u"	"oan"	"ob."
"of."	"ofc"	"ofm"	"ofn"	"ofs"	"ogl"	"oj."	"om."	"omm"	"on."	"ons"	"oog"	"org"	"orp"	"oun"	"ov."	"oz."	"q.c"	"r.u"	"rch"
"rg."	"rls"	"rp."	"rs."	"rth"	"ru."	"rug"	"s.c"	"s.o"	"s.u"	"sci"	"sco"	"sea"	"sec"	"sex"	"sfa"	"sin"	"s-l"	"sn."	"sp-"
"ss-"	"sta"	"ste"	"sys"	"t.u"	"tat"	"tec"	"teo"	"ter"	"tk."	"tla"	"to."	"ty."	"tyn"	"tyo"	"u.o"	"ud."	"uen"	"uh."	"uj."
"uni"	"unt"	"us."	"uv."	"uw."	"v.c"	"vwb"	"w.c"	"web"	"wes"	"wnl"	"ws."	"ww."	"www"	"y.u"	"yah"	"yof"	"you"	"z.c"	"-lo"

le HTTPS, etc.). Nous avons donc choisi de vérifier que l'URL affichée ne présente pas d'indication de port autre que le port standard.

Si on détecte l'indication d'un port alternatif pour le protocole affiché, nous attribuons un score négatif à l'heuristique correspondant. En l'absence de numéro de port, le score est nul. Cet heuristique s'applique à l'analyse du FQDN exclusivement (cf. tableau 3.3).

La phase d'étalonnage démontre que ce cas de figure n'a jamais été rencontré sur whitelist (cf. tableau 3.4).

Des exemples d'URL contrefaites qui utilisent un numéro de port non-standard : `http://raceobject.ru:8080/index.php?pid=14/` et `http://hillchart.com:8080/index.php?pid=14/` emploient le port 8080 en association avec le protocole HTTP. Ce port 8080 est souvent utilisé pour faire fonctionner un serveur web alternatif, lorsque l'utilisateur ne possède pas les droits administrateurs[1] (c.-à-d. qu'il n'a pas l'autorisation d'utiliser les ports *well-known* 0 à 1023) ou que le port 80 de la machine est déjà utilisé par un autre serveur web.

3.4.1.2 Catégorie Triplets et mots-clés (dits de phishing)

3.4.1.2.1 Heuristiques N° 6 et 7 : Nombre de triplets dans l'URL Une étude préalable, réalisée conjointement avec le FAI Orange, menée sur des whitelists et des blacklists (comprenant au cumulé jusqu'à 56 227 URLs ou composants d'URLs) nous a permis d'identifier un certain nombre de triplets de caractères souvent présents dans les URLs de phishing.

Cette étude a été réalisée à partir du logiciel *Khiops* proposé par Boullé [Bou], qui offre un outil de préparation et de modélisation des données, utilisable à des fins d'exploration de bases de données (p.ex. pour des études d'usage) [Bou08]. Il permet d'aboutir à une analyse prédictive des variables d'entrées (c.-à-d. les composants d'URLs) via une approche statistique de type Bayésienne.

Des résultats de cette étude, nous avons extrait un ensemble de 240 triplets de caractères - considérés par l'outil comme les plus présents dans les URLs de phishing -, que nous utilisons comme base de notre analyse de l'URL visitée (cf. tableau 3.2).

Parmi ces triplets, nous avons constaté que 4 d'entre eux étaient très fréquents dans tous FQDNs (y compris ceux des URLs légitimes) : `www`, `ww.`, `.co` et `com`. Il peut alors apparaître surprenant que l'outil nous les énonce comme caractéristiques d'une URL de phishing. Ceci s'explique néanmoins par le fait que ces triplets - habituellement réservés à une utilisation au sein du FQDN (par les URLs légitimes) - se retrouvent fréquemment utilisés tant dans le FQDN que dans l'arborescence des URLs de phishing (via le rappel du FQDN légitime, cf. section 2.2.1).

Nous avons donc assigné les scores des heuristiques en conséquence :

– *Nombre de triplets dans le FQDN :* la présence de 4 triplets est choisie comme valeur charnière de la décision de légitimité d'un site, afin de correspondre aux 4 triplets courants cités précédemment (cf. tableau 3.3). Jusqu'à cette valeur, le score de l'heuristique est positif ou nul. Au-delà, il est négatif.

  L'étalonnage réalisé fait apparaître que les triplets sont nombreux, particulièrement dans les FQDNs de phishing. Toutefois, il semble important de préserver les URLs légitimes en deçà de 3

1. C'est le cas typique d'un site de phishing hébergé et mis en ligne au travers d'un réseau de botnets, cf. section 2.3.

Tableau 3.3 – Valeurs charnières pour les seuils de décision des heuristiques implémentés

N°	Heuristique	Valeur charnière
1	Nombre de points (.) dans l'URL	3
2	Nombre d'arobas (@) dans l'URL	1
3	Nombre de double slash (//) dans l'URL	1 au delà des (//) énoncés après le protocole (p.ex. `http://`)
4	Présence d'une adresse IP dans le FQDN	1
5	Présence d'un numéro de port dans le FQDN	1, si non standard
6	Nombre de triplets dans le FQDN	4
7	Nombre de triplets dans l'arborescence de l'URL (hors FQDN)	1
8	Nombre de mots-clés dans l'arborescence de l'URL (hors FQDN)	1
9	Présence d'un TLD "sensible" dans le FQDN	si appartenance au Groupe 1 ou au Groupe 2
10	Présence d'un TLD "sensible" dans l'arborescence de l'URL (hors FQDN)	si appartenance au Groupe 1 ou au Groupe 2
11	Comparaison entre TLD et pays hébergeur du site	si non correspondance
12	Evaluation de la balise de Titre	si absence du domaine/FQDN
13	Evaluation de balise Formulaire	si absence du domaine/FQDN
14	Evaluation de lien Image	si absence du domaine/FQDN
15	Evaluation d'autre lien	si absence du domaine/FQDN
16	HTTPS et zones de login	si présence de champs *PASSWORD* ou *LOGIN* hors connexion HTTPS
17	Evaluation de balise de Description Meta	si absence du domaine/FQDN
18	Evaluation de balise de mots-clés Meta	si absence du domaine/FQDN
19	Evaluation de balise Script	si absence du domaine/FQDN
20	Evaluation de balise Link	si absence du domaine/FQDN

triplets (cf. tableau 3.4).
- *Nombre de triplets dans l'arborescence de l'URL (hors FQDN) :* au-delà d'1 triplet (cf. tableau 3.3), le score assigné est négatif. Sinon, le score attribué est positif.
  L'étalonnage démontre qu'un nombre conséquent d'URLs de phishing utilise les triplets à de multiples reprises dans l'arborescence de l'URL (cf. tableau 3.4).

Des exemples d'occurrences de triplets :
- L'URL légitime `http://bluewin.ch/` utilise 2 triplets dans son FQDN et 0 triplet dans le reste de l'URL.
- L'URL légitime `http://www.cashfiesta.com/php/login.php` utilise 7 triplets dans son FQDN et 0 triplet dans le reste de l'URL.
- L'URL contrefaite `http://www.agb.com.ve/maps/160703/survey/login.htm?paypal.com/us/cgi-bin/webscr?cmd=_login-submit&dispatch=5885d80a13c0db1f1ff80d546411d7f8a8350c132bc41e0934cfc023d4e8f9e5` utilise 6 triplets dans son FQDN et 4 triplets dans le reste de l'URL.
- L'URL contrefaite `http://www.jacuzzifilms.com/plugins/editors-xtd/modulesanywhere/css/wellsfargo-online.php` utilise 8 triplets dans son FQDN et 1 triplet dans le reste de l'URL.

3.4.1.2.2 Heuristique N° 8 : Nombre de mots-clés dans l'arborescence de l'URL (hors FQDN) Après une étude approfondie des URLs de phishing, à l'image des travaux menés par Garera et al. [GPCR07], nous avons sélectionné 5 mot-clés (c.-à-d. `www`, `http`, `login`, `logon`, `paypal`) qui apparaissent régulièrement dans les URLs de phishing. Le choix des 4 premiers mots-clés s'explique par le fait que les URLs de phishing tendent à usurper des pages de login et/ou mentionner le FQDN légitime dans l'arborescence de l'URL pour leurrer les Internautes. Le choix du 5ème mot-clé se justifie quant à lui par le fait que Paypal apparaît comme une cible privilégiée des attaquants.

Toutefois, nous notons que ces mots-clés se retrouvent également fréquemment dans les FQDNs des

URLs légitimes. Nous choisissons donc de limiter l'analyse de cet heuristique à l'arborescence de l'URL (hors FQDN).

Ainsi, dès l'apparition d'un mot-clé (cf. tableau 3.3), le score attribué est négatif. En leur absence, le score attribué est positif.

La phase d'étalonnage laisse apparaître qu'il est assez fréquent de trouver l'un des mots-clés (typiquement `login` ou `logon`) dans le nom de fichier ou les données complémentaires précisées dans les URLs de login légitimes (cf. tableau 3.4).

Des exemples d'occurrences de mots-clés dans les URLs :
- L'URL légitime `https://cas.it-sudparis.eu/cas/login?service=http://gaspar.it-sudparis.eu/uPortal/Login` utilise 2 fois le mot-clé `login` dans son arborescence.
- L'URL légitime `https://login.yahoo.com/config/mail?&.src=ym&.intl=fr` utilise le mot-clé `login` dans son FQDN uniquement.
- L'URL contrefaite `http://www.neural-net.ca/store/images/webscr/1/www.paypal.co.uk/` utilise les mots-clés `www` et `paypal` dans son arborescence.

3.4.1.3 Catégorie TLD

3.4.1.3.1 Heuristiques N° 9 et 10 : Présence d'un TLD "sensible" dans l'URL Parmi les triplets retournés par notre analyse (cf. section 3.4.1.2), nous avons remarqué que certains d'entre eux correspondent à des TLDs. Par ailleurs, chaque trimestre, l'APWG publie un Top 10 des pays hébergeurs des sites de phishing (en indiquant le proportion de sites de phishing qu'ils hébergent).

Nous avons donc décider d'utiliser cette information, pour pénaliser les URLs dont le TLD est issu d'un pays connu pour sa proportion à héberger les sites de phishing. Pour ce faire, nous avons créé 2 groupes de TLDs : le Groupe 1 contient les TLDs des pays qui hébergent en moyenne plus de 40% de sites de phishing, tandis que le Groupe 2 inclut le reste des TLDs connus pour l'hébergement de sites contrefaits. Plus précisément, nous avons recensé l'ensemble des TLDs signalés par l'APWG sur une période de 12 mois. Pour chacun d'entre eux, nous avons calculé la moyenne d'hébergement sur la période étudiée. Ainsi, nous pouvons mieux faire face à l'éventuelle mouvance des pays hébergeurs.

Nous attribuons alors un score négatif aux URLs qui utilisent un TLD des Groupes 1 et 2 (ce score est doublé pour les URLs dont le TLD appartient au Groupe 1), tandis que le score assigné est nul pour les URLs en provenance d'autres TLDs (cf. tableau 3.3). Cet heuristique s'applique à l'intégralité de l'URL, afin de tenir compte d'éventuelles redirections de sites au sein des URLs de phishing.

La phase d'étalonnage démontre que les seuils optimum sont identiques sur blacklist et whitelist (cf. tableau 3.4).

A ce jour, les pays hébergeurs de sites de phishing pénalisés par notre moteur de détection sont :
- Groupe 1 : US et UM (tous deux correspondent aux États-Unis).
- Groupe 2 : SE, CN, CA, UK, GB, DE, KP, FR, PM, RE, TF, WF, GT, RU, SU, AN, NL, TW, RO, PL, ES, HU, HK et BR.

3.4.1.3.2 Heuristique N° 11 : Comparaison entre TLD et pays hébergeur du site A partir de l'étude de McGrath et al. [MG08] qui indique que la plupart des sites de phishing ne sont pas hébergés dans le pays annoncé par leur TLD, nous avons cherché à évaluer la corrélation entre TLD annoncé par l'URL et pays hébergeur du site.

Pour ce faire, nous nous sommes appuyés sur un plug-in additionnel existant pour Mozilla Firefox : *World IP* [wor]. Celui-ci permet de donner diverses informations sur le site web visité (p.ex. adresse IP, pays hébergeur, etc. - cf. figure 3.2 - grâce à des requêtes reverse DNS, ping, traceroute, etc.)[1].

Néanmoins, parce que de nombreux sites légitimes ne sont pas non plus hébergés dans le pays indiqué par le TLD - typiquement dans le cas de multinationales et/ou de TLD génériques -, nous avons choisi d'utiliser ce critère de manière favorable exclusivement (cf. tableau 3.3).

1. D'autres plug-ins de même type sont disponibles pour Mozilla Firefox, mais *World IP* est celui qui nous est apparu le plus complet/moins restrictif en terme d'informations délivrées.

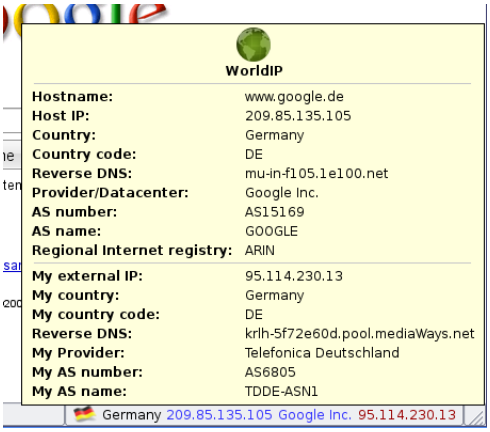


FIGURE 3.2 – Aperçu des informations délivrées par World IP [wor]

Ainsi, si le TLD du FQDN contenu dans l'URL correspond au code pays (champ *Country-Code*) annoncé par *World IP*, le score assigné est positif. Sinon, il est nul (cf. tableau 3.4).

Des exemples d'URLs et pays hébergeurs associés :
- Les sites légitimes `https://particuliers.societegenerale.fr/` et `http://www.013netvision.net.il/` sont hébergés dans le pays indiqué par leur TLD.
- Les URLs légitimes `http://www.dell.fr` et contrefaites `http://estiloforrozeiro.com.br/paypal.co.uk/_cmd-check.php?cmd_check=logged&SESSION-ID=8859-1&uid=F654557d45Af4t` et `http://secure.paypal.fr.freezonesuae.com/login/`, sont toutes trois hébergées aux États-Unis, bien que cela ne transparaisse pas au niveau de leurs TLDs.

3.4.2 Code source HTML

Au-delà de l'étude de l'URL, il peut être intéressant d'examiner le contenu de la page web (autrement dit, le code source HTML de la page) pour détecter les sites de phishing.

En effet, notre étude approfondie des contenus des pages webs laisse apparaître que les sites de phishing présentent un certain nombre d'incohérences, entre code source de la page et URL/FQDN (cf. section 2.2.2). Pour cette raison, nous nous intéressons ici à définir/utiliser des heuristiques qui visent à comparer ces informations, via l'étude de la cohérence des contenus de balises HTML vs. le FQDN/domaine visité.

Ces heuristiques sont regroupés en 3 catégories : *Code source HTML*, *Page de Login* et *Autres balises HTML* (cf. tableau 3.1).

3.4.2.1 Catégorie Code source HTML

3.4.2.1.1 Heuristique N° 12 : Evaluation de la balise de Titre

La balise <TITLE> présente dans le document HTML est généralement utilisée pour spécifier le titre du document.

Dans les URLs légitimes, on y retrouve principalement une information en lien avec le domaine/FQDN visité. Des exemples de contenu de balises <TITLE> de sites légitimes :
- Dans la page de login `http://www.facebook.com/`, on trouve le titre `Bienvenue sur Facebook. Connectez-vous, inscrivez-vous ou découvrez !` qui contient le nom de domaine.
- Dans la page d'accueil `http://www.google.fr/`, on trouve le titre `Google` qui est le nom de domaine et du service.
- Dans la page de login Yahoo ! `https://login.yahoo.com/config/mail?&.src=ym&.intl=fr`, on trouve le titre `Ouverture de session Yahoo! - Création de compte` qui contient le nom de domaine.

TABLEAU 3.4 – Seuils de décision optimum des heuristiques pour la Whitelist et la Blacklist

N°	Heuristique	Seuils optimum pour la Whitelist	Seuils optimum pour la Blacklist
1	Nombre de points (.) dans l'URL	-2 si quantité (.) > 10 -1 si quantité (.) > 5 0 si quantité (.) ≤ 3 1 si quantité (.) = 1	-2 si quantité (.) > 10 -1 si quantité (.) > 5 0 si quantité (.) ≤ 3 1 si quantité (.) = 1
2	Nombre d'arobas (@) dans l'URL	-1 si quantité (@) ≥ 1 1 si quantité (@) = 0	-1 si quantité (@) ≥ 1 1 si quantité (@) = 0
3	Nombre de double slash (//) dans l'URL	-1 si quantité (//) ≥ 1 1 si quantité (//) = 0	-2 si quantité (//) ≥ 2 -1 si quantité (//) ≥ 1 1 si quantité (//) = 0
4	Présence d'une adresse IP dans le FQDN	-2 si adresse IP 0 si pas adresse IP	-2 si adresse IP 0 si pas adresse IP
5	Présence d'un numéro de port dans le FQDN	*Aucune URL concernée*	-1 si non standard 0 si pas de numéro de port
6	Nombre de triplets dans le FQDN	-2 si quantité triplets > 10 -1 si quantité triplets ≤ 10 0 si quantité triplets ≤ 4, ou pas de FQDN[1] 1 si quantité triplets ≤ 2 2 si quantité triplets = 0	-2 si quantité triplets > 10 -1 si quantité triplets ≤ 10 0 si quantité triplets ≤ 4, ou pas de FQDN[1] 1 si quantité triplets = 0
7	Nombre de triplets dans l'arborescence de l'URL (hors FQDN)	-1 si quantité triplets ≥ 1 1 si quantité triplets = 0	-2 si quantité triplets > 4 -1 si quantité triplets ≥ 2 0 si quantité triplets = 1 1 si quantité triplets = 0
8	Nombre de mots-clés dans l'arborescence de l'URL (hors FQDN)	-3 si quantité mots-clés > 4 -2 si quantité mots-clés > 2 -1 si quantité mots-clés = 2 0 si quantité mots-clés = 1 1 si quantité mots-clés = 0	-3 si quantité mots-clés > 4 -2 si quantité mots-clés ≥ 2 -1 si quantité mots-clés = 1 1 si quantité mots-clés = 0
9	Présence d'un TLD "sensible" dans le FQDN	-2 si TLD ∈ Groupe 1[2]	-2 si TLD ∈ Groupe 1[2]
10	Présence d'un TLD "sensible" dans l'arborescence de l'URL (hors FQDN)	-1 si TLD ∈ Groupe 2[2] 0 si TLD ∈ aucun Groupe	-1 si TLD ∈ Groupe 2[2] 0 si TLD ∈ aucun Groupe
11	Comparaison entre TLD et pays hébergeur du site	0 si pas correspondance 1 si correspondance	*Pas nécessaire*[3]
12	Evaluation de la balise de Titre	-1 si ∅ ou pas correspondance avec FQDN/domaine 2 si correspondance avec FQDN/domaine	-2 si pas correspondance avec FQDN/domaine -1 si ∅ 1 si correspondance avec FQDN/domaine
13	Evaluation de balise Formulaire		
14	Evaluation de lien Image		
15	Evaluation d'autre lien		
16	HTTPS et zones de login	-1 si zone de login hors HTTPS 0 si pas de zone de login 3 si zone de login en HTTPS	-3 si zone de login hors HTTPS 0 si pas de zone de login 2 si zone de login en HTTPS
17	Evaluation de balise de Description Meta	0 si ∅ ou pas de correspondance avec FQDN/domaine 1 si correspondance avec FQDN/domaine	-1 si ∅ ou pas de correspondance avec FQDN/domaine 1 si correspondance avec FQDN/domaine
18	Evaluation de balise de mots-clés Meta		
19	Evaluation de balise Script		
20	Evaluation de balise Link		

[1] c.-à-d. remplacé par une adresse IP.
[2] cf. section 3.4.1.
[3] car pas de pénalité négative possible, cf. section 3.4.1.

A contrario dans les URLs de phishing, on trouve rarement cette corrélation. En effet de par les techniques d'aspiration de sites webs utilisées par les attaquants pour créer leur contrefaçon et/ou leur volonté de minimiser les modifications apportées, on retrouve très souvent le contenu de la balise <TITLE> du site légitime dans la page web contrefaite. Des exemples de contenus de balises <TITLE> de sites contrefaits :

- Dans la page contrefaite de login `http://61.146.118.23:8088/a/`, on trouve le titre `Welcome to eBay - Sign in - Error` qui contient le nom de domaine usurpé `eBay`.
- Dans la page contrefaite de login `http://buyacaravan.net/webscr.php`, on trouve le titre `Send Money, Pay Online, and Receive Money - all with PayPal` qui contient le nom de domaine usurpé `Paypal`.
- Dans la page contrefaite de login `http://service.eshost.es/index3.htm`, on retrouve le titre `¡Bienvenido a Facebook en Español!` qui contient le nom de domaine usurpé `Facebook`.

Parce qu'une URL de phishing - bien qu'elle soit parfois très ressemblante - n'utilise pas (ou qu'exceptionnellement) le domaine légitime dans son FQDN, nous avons considéré qu'il pouvait être intéressant de vérifier la cohérence de ces informations (c.-à-d. FQDN/domaine visité et contenu de la balise <TITLE>).

Ainsi, s'il y a correspondance, le score assigné est positif. A contrario, s'il n'y a pas correspondance ou si la balise est vide, le score attribué est négatif (cf. tableau 3.3).

Notre phase d'étalonnage démontre que l'étude de cet heuristique est particulièrement déterminante pour les sites de phishing existants actuellement. Le score positif profite quasi-exclusivement aux sites légitimes. Une balise vide est souvent (mais pas forcément) révélatrice d'une contrefaçon. Enfin, il est encore plus fréquent qu'une incohérence entre FQDN et <TITLE> soit révélatrice d'une contrefaçon (cf. tableau 3.4).

3.4.2.1.2 Heuristiques N° 13, 14 et 15 : Evaluation de balises de formulaire (FORM) et de liens (IMG et A HREF) Une contrefaçon de site web se doit de conserver un maximum de contenu du site légitime, y compris les liens accessibles depuis la page usurpée. En effet, si un utilisateur souhaite consulter une information d'aide (avant de s'authentifier par exemple), et qu'il se trouve face à une page d'erreur ou une page approximative, son attention risque d'être éveillée. Or la qualité première d'une contrefaçon est de s'approprier le décor original au maximum. Quelle meilleure technique que de conserver toutes les redirections d'origine (autres que celles de login par exemple) pour leurrer l'utilisateur ? Non seulement, cela permet de rendre l'attaque plus transparente (visuellement parlant dans l'affichage rendu par le navigateur) mais en plus, cela minimise les efforts de l'attaquant qui peut se contenter de contrefaire la première page visitée par l'Internaute.

Le revers de la médaille, c'est que ce type de technique n'est pas du tout transparente dans le code source HTML. En effet, quasiment tous les liens spécifiés font référence au FQDN/domaine légitime et non au FQDN visité, comme Chen et al. [CG06] l'ont déjà remarqué.

De cet état de fait, nous avons choisi de déduire des tests heuristiques. Ainsi, nous nous intéressons aux balises les plus révélatrices de ces incohérences et/ou les plus soumises à contrefaçon : <IMG>, <A HREF> et <FORM>. En effet, toutes 3 contiennent généralement des URLs et/ou des paramètres additionnels qui utilisent le domaine légitime (cf. section 2.2.2 pour plus de détails sur ces balises).

Pour ces trois heuristiques, si le domaine visité apparaît dans le contenu de la balise, le score assigné est positif. A contrario, s'il n'y a pas correspondance ou si la balise est vide, le score attribué est négatif (cf. tableau 3.3).

Précisons que dans la version actuelle de *Phishark*, notre analyse se cantonne à l'étude de la première balise de chaque type trouvée. Dans une version ultérieure nous envisageons d'améliorer la robustesse de notre solution en étendant cet heuristique à une analyse de l'ensemble des balises de chaque type (p.ex. chaque score d'heuristique serait alors déterminé par l'intermédiaire d'une moyenne).

Notre phase d'étalonnage démontre que le score positif profite quasi-exclusivement aux sites légitimes, qu'une balise vide est souvent (mais non forcément) révélatrice d'une contrefaçon, et qu'il est encore plus fréquent qu'une incohérence entre FQDN et une balise soit révélatrice d'une contrefaçon (cf. tableau 3.4).

Des exemples des contenus de ces 3 balises :
- Le site légitime de météo italien Yahoo !`http://it.meteo.yahoo.com/Europa/italia/` présente les contenus suivants :
  - La balise <FORM> contient `action="http://it.search.yahoo.com/search"`,
  - La balise <IMG> contient `src="http://row.bc.yahoo.com/b?P=0WwM8Ff4erSlvqiQTchZWhi"`
  - La balise <A HREF> contient `"http://it.notizie.yahoo.com/"`

  Dans chacune des trois balises, nous retrouvons le domaine légitime.
- Le site contrefait `http://permitds.com/system/` qui usurpe le site web légitime USAA `https://www.usaa.com/`...présente les contenus suivants :
  - La balise <FORM> contient `action="done4.php"method="post"onSubmit="returnvg(this,'id,pw,q1,a1,q2,a2,q3,a3,fn,cc,sc,pin,em','Pleasenoteallfieldsarerequired.');"`,
  - La balise <IMG> contient `src="usaa_log.png"width="53"height="55"`
  - La balise <A HREF> contient `"https://www.usaa.com/inet/ent_home/CpHome"class="logo""`

  Aucune balise ne désigne explicitement le domaine visité. A contrario, deux d'entre elles font référence au domaine légitime.

3.4.2.2 Catégorie Page de Login

3.4.2.2.1 Heuristique N° 16 : HTTPS et zones de login Les sites de phishing visent à usurper des sites de login légitimes. Ces derniers sont accédés - en grande majorité - au travers d'une URL qui utilise le protocole HTTPS afin d'établir une connexion sécurisée. Précisons toutefois que des exceptions existent : on peut notamment citer le cas de `Facebook` qui sécurise l'envoi des données de façon totalement transparente pour l'utilisateur, ou de sites d'abonnement divers et variés (p.ex. des abonnements à des journaux, etc.). Néanmoins, à ce jour, nous n'avons pas rencontré de site de login qui donne un accès direct à des sources financières (carte ou compte bancaire) sans un affichage explicite en HTTPS.

A contrario les contrefaçons des sites légitimes sont en grande majorité accédées via le protocole HTTP. Il est en effet plus difficile et/ou moins souhaitable pour l'attaquant de disposer d'un certificat valide pour son site (cf. section 2.2.1). Néanmoins, le manque de vigilance des utilisateurs combinée à l'utilisation de subterfuges (tels que l'affichage du cadenas habituellement affiché par une connexion HTTPS, l'insertion de logos de sécurité, etc.) préservent l'efficacité de leurs attaques (cf. section 2.2.2).

De ce constat, nous avons imaginé qu'il serait intéressant de développer un heuristique qui s'attache à vérifier qu'une page web qui présente une zone de login, le fasse sous couvert de la mention explicite d'une connexion sécurisée en HTTPS. Tout en ayant conscience que de nombreux sites de login (moins sensibles que les sites d'accès direct aux finances) risquent de générer des faux-positifs.

Pour ce faire, nous parcourons le code source HTML de la page, à la recherche d'une zone d'authentification mentionnée par les champs *TYPE=PASSWORD* ou *TYPE=LOGIN*, typiquement trouvés au sein des balises <FORM> et <INPUT> pour désigner des zones de saisie utilisateur. Puis, nous examinons l'URL afin de voir si elle utilise le protocole HTTPS (cf. tableau 3.3). Si c'est le cas, le score assigné est positif. Sinon, il est négatif. Enfin, en l'absence de champs *TYPE=PASSWORD* ou *TYPE=LOGIN*, le score est nul.

Notre phase d'étalonnage démontre qu'il faut fortement privilégier les sites qui utilisent une connexion HTTPS. A contrario, de par le nombre de sites légitimes qui présentent des zones de login sous un affichage HTTP, il apparaît plus problématique de pondérer aussi fortement l'absence de HTTPS (cf. tableau 3.4).

Des exemples de sites légitimes qui utilisent - ou non - une connexion HTTPS :
- Mention explicite d'une connexion HTTPS : les pages de login des sites bancaires `https://www.paypal.com/`, `https://particuliers.societegenerale.fr/`, `https://www4.bankofamerica.com/hub/index.action?template=signin`, des sites marchands `https://www.amazon.fr/gp/css/ho mepage.html?ie=UTF8&ref_=topnav_ya`, etc.
- Zone d'authentification, hors HTTPS : les pages d'accueil des réseaux sociaux `http://www.facebook.com/`, `http://www.linkedin.com/`, des sites d'informations `http://www.lemonde.fr/`, des sites marchands `http://www.pixmania.com/fr/fr/c_action/mon_compte/index.html`, etc.

3.4.2.3 Catégorie Autres balises HTML

Nous avons constaté que les analyses des balises <IMG>, <FORM> et <A HREF> comparées au FQDN/domaine visité sont particulièrement efficaces pour identifier les sites de phishing. Nous avons donc souhaité étendre ces heuristiques à d'autres balises. Nous nous sommes alors intéressés aux balises <META DESCRIPTION>, <META KEYWORDS>, <SCRIPT> et <LINK>.

Précisons que les balises <META> sont nombreuses (description, keywords, author, date, etc.) mais que leur utilisation n'est pas obligatoire. Comme tous champs optionnels, il apparaît hasardeux de les utiliser en tant que critères de décision. Néanmoins, nous en avons retenu deux : la balise <META DESCRIPTION> qui est présente dans la quasi-totalité des sites que nous avons étudiés, et la balise <META KEYWORDS> que nous trouvons particulièrement intéressante puisqu'elle est principalement utilisée par les moteurs de référencement (un attaquant aura donc peu d'intérêt à la modifier, mieux vaudrait encore la supprimer que d'y mettre des informations erronées).

3.4.2.3.1 Heuristique N° 17 : Evaluation de balise de Description META La balise <META DESCRIPTION> permet de donner une description du site. Dans les pages légitimes, on y retrouve donc souvent la mention du domaine visité. Dans les pages de phishing, le contenu de cette balise est souvent inchangé par rapport au site original (c.-à-d. il contient le domaine légitime), ou la balise est tout simplement supprimée.

De par sa présence quasi-permanente dans les sites légitimes, nous lui accordons un traitement particulier (vs. les balises des heuristiques N° 18, 19 et 20 détaillés ci-après). Ainsi, si nous retrouvons le domaine visité dans la balise, le score attribué est positif. Si la balise est inexistante, le score est nul. Enfin, si la balise est incohérente avec le domaine visité, le score assigné est négatif (cf. tableau 3.3).

Notre phase d'étalonnage démontre qu'il est également possible de trouver des incohérences sur sites légitimes. Néanmoins, ces cas sont surtout l'apanage des sites contrefaits, de même que l'inexistence de la balise (cf. tableau 3.4).

Des exemples de balises <META DESCRIPTION> :

- La page d'accueil légitime Facebook `http://www.facebook.com/` contient la description `"Facebook est un réseau social qui vous relie à des amis, des collègues de travail, des camarades de classe..."`. On y retrouve aisément la mention du domaine.
- La page de login légitime de la banque BNP PARIBAS `https://www.secure.bnpparibas.net/banque/portail/particulier/HomeConnexion?`...contient la description `"Tous les produits et services de votre banque en France Accéder à vos comptes, titres et Bourse."`. On n'y retrouve aucune mention du domaine.
- La page contrefaite `http://permitds.com/system/` qui usurpe le site web légitime USAA `https://www.usaa.com/`...ne comporte pas de description.
- La page contrefaite `http://www.lospws2.co/ca/` qui usurpe la page web légitime Paypal `https://www.paypal.com/us/cgi-bin/webscr?cmd=_home&locale.x=en_US`...contient la description `"PayPal lets you send money to anyone with email. PayPal is free for consumers"`.... On y retrouve mention du domaine légitime et non du domaine visité.

3.4.2.3.2 Heuristiques N° 18, 19 et 20 : Evaluation de balises de mots-clés META, SCRIPT et LINK Les balises <META KEYWORDS>, <SCRIPT> et <LINK> (cf. section 2.2.2 pour plus de détails) sont d'usage moins courant que les balises étudiées dans les heuristiques vus précédemment. De plus, elles ne contiennent pas forcément une référence au domaine en se limitant parfois à une arborescence ou un nom de fichier appelé.

Il apparaît donc difficile de pénaliser d'éventuelles incohérences entre domaine visité et contenu de la balise. Nous réservons donc ces heuristiques à un usage bénéfique exclusivement, ceci afin de privilégier les sites légitimes.

Ainsi, si nous retrouvons le domaine visité dans la balise, le score attribué est positif. Si la balise est inexistante ou incohérente avec le domaine, le score est nul (cf. tableau 3.3).

Notre phase d'étalonnage démontre qu'il est important de ne pas pénaliser les incohérences pour les sites légitimes, contrairement au cas des sites contrefaits (cf. tableau 3.4).

TABLEAU 3.5 – Seuils de décision retenus après fusion des seuils optimum sur Whitelist et Blacklist

N°	Heuristique	Seuils de décision retenus
1	Nombre de points (.) dans l'URL	-2 si quantité (.) > 10 -1 si quantité (.) > 5 0 si quantité (.) ≤ 3 1 si quantité (.) = 1
2	Nombre d'arobas (@) dans l'URL	-1 si quantité (@) ≥ 1 1 si quantité (@) = 0
3	Nombre de double slash (//) dans l'URL	-1 si quantité (//) ≥ 1 et redirection vers un domaine différent 1 si quantité (//) ≥ 1 et redirection vers le même domaine 1 si quantité (//) = 0
4	Présence d'une adresse IP dans le FQDN	-2 si adresse IP 0 si pas adresse IP
5	Présence d'un numéro de port dans le FQDN	-1 si non standard 0 si pas de numéro de port
6	Nombre de triplets dans le FQDN	-2 si quantité triplets > 10 -1 si quantité triplets ≤ 10 0 si quantité triplets ≤ 4, ou pas de FQDN[1] 1 si quantité triplets ≤ 2 2 si quantité triplets = 0
7	Nombre de triplets dans l'arborescence de l'URL (hors FQDN)	-2 si quantité triplets > 4 -1 si quantité triplets ≥ 1 1 si quantité triplets = 0
8	Nombre de mots-clés dans l'arborescence de l'URL (hors FQDN)	-3 si quantité mots-clés > 4 -2 si quantité mots-clés ≥ 2 -1 si quantité mots-clés = 1 1 si quantité mots-clés = 0
9	Présence d'un TLD "sensible" dans le FQDN	-2 si TLD ∈ Groupe 1[2]
10	Présence d'un TLD "sensible" dans l'arborescence de l'URL (hors FQDN)	-1 si TLD ∈ Groupe 2[2] 0 si TLD ∈ aucun Groupe
11	Comparaison entre TLD et pays hébergeur du site	0 si pas correspondance 1 si correspondance
12	Evaluation de la balise de Titre	-2 si pas correspondance avec FQDN/domaine -1 si ∅ 2 si correspondance avec FQDN/domaine
13	Evaluation de balise de Formulaire	-1 si ∅ ou pas correspondance avec FQDN/domaine
14	Evaluation de lien Image	1 si correspondance avec FQDN/domaine
15	Evaluation d'autre lien	
16	HTTPS et zones de login	-2 si zone de login hors HTTPS 0 si pas de zone de login 3 si zone de login en HTTPS
17	Evaluation de balise de description Meta	-1 si pas de correspondance avec FQDN/domaine 0 si ∅ 1 si correspondance avec FQDN/domaine
18	Evaluation de balise de mots-clés Meta	0 si ∅ ou pas de correspondance avec FQDN/domaine
19	Evaluation de balise Script	1 si correspondance avec FQDN/domaine
20	Evaluation de balise Link	

[1] c.-à-d. remplacé par une adresse IP.
[2] cf. section 3.4.1.

Des exemples de contenus de ces 3 balises :

- La page de login légitime de la banque BNP PARIBAS `https://www.secure.bnpparibas.net/banque/portail/particulier/HomeConnexion?`...contient `"BNPPARIBAS.NET, connexion, comptes,titres,bourse"` pour les <META KEYWORDS>. Les balises <SCRIPT> et <LINK> se limitent quant à elles à des arborescences et noms de fichiers (p.ex. `src="/banque/PA_CanalnetApp/scripts/canalnetForms.js"`).
- La page contrefaite `http://www.lospws2.co/ca/` qui usurpe la page web légitime Paypal `https://www.paypal.com/us/cgi-bin/webscr?cmd=_home&locale.x=en_US`...ne contient pas <META-KEYWORDS>. Les balises <SCRIPT> et <LINK> se limitent à des arborescences et noms de fichiers.
- La page légitime MSN `http://www.msn.com/?st=1` contient le lien `src="http://col.stj.s-msn.com/br/sc/js/jquery/jquery-1.4.2.min.js"` où figure le nom de domaine.

Pour conclure notre description des heuristiques implémentés, le tableau 3.5 expose les seuils décisionnels actuellement implémentés par *Phishark*. Ils ont été choisis pour les raisons exposées précédemment à partir des seuils optimum relevés sur blacklist et whitelist.

3.4.3 Description de *Phishark*

Le moteur de détection de notre barre d'outils a été développé en Javascript. L'interface graphique a quant à elle été développée en langage XUL (pour *Xml-based User interface Language*) - basé sur le XML (pour *eXtensible Markup Language*) -. Ainsi, chaque action/événement de l'interface visuelle fait appel à une fonction Javascript. Le choix de ces langages orientés client réside essentiellement dans leur légèreté, simplicité, compatibilité et l'absence de besoin de compilation. L'interface graphique de la barre d'outils (p.ex. les indications de localisation des images, sa taille, etc.) est définie et structurée par une feuille de style CSS (pour *Cascading Style Sheet*).

Basés sur des travaux précédents [ECH08] [WMG06] menés sur diverses barres d'outils anti-phishing et notre analyse de leurs différentes interfaces visuelles, nous avons choisi de développer une barre d'outils relativement simple, visuellement aussi explicite que possible (via des jeux de couleurs ou de changements d'icônes), principalement axée autour de messages de notifications actifs (p.ex. via des pop-up). En effet, ces travaux précédents laissent apparaître que les notifications actives sont plus efficaces auprès des utilisateurs, car elles nécessitent obligatoirement une action de l'Internaute en cas de site suspect.

L'intégration générale de la barre *Phishark* au sein du navigateur est illustrée en figure 3.3. Par défaut, la barre d'outils se positionne tel qu'indiqué sur la figure. On remarque également une zone d'informations associée à l'utilisation de *World IP*[1] en bas à droite du navigateur. En amenant la souris sur une des adresses IP (représentatives du côté client ou du côté serveur web) affichée par *World IP*, une vue plus détaillée des informations délivrées par ce plug-in est dévoilée.

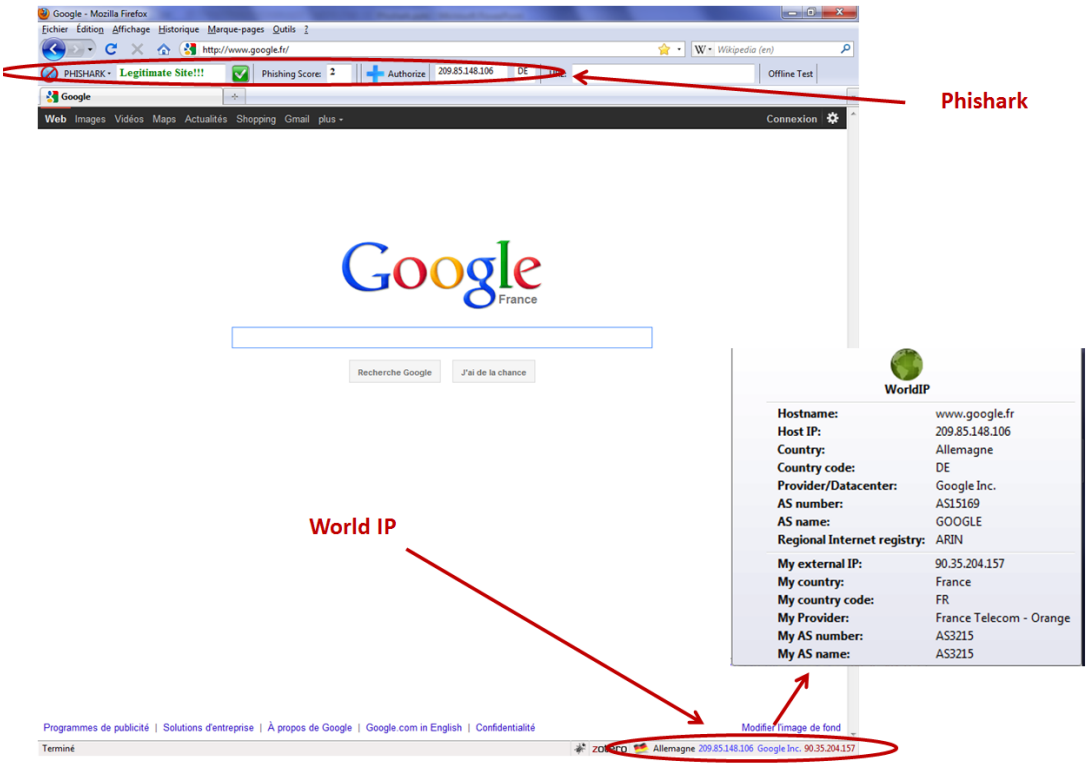


FIGURE 3.3 – Schéma d'intégration de Phishark dans le navigateur web Mozilla Firefox, incluant le plug-in World IP [wor]

La figure 3.4 zoome sur *Phishark*. On y voit que la barre est constituée de 5 zones :

– la zone 1 donne accès au menu de configuration de la barre pour : définir/modifier une white-list personnelle, redéfinir le seuil de notification *Risqué* (par défaut établi à 0, mais une plage

1. Pour plus d'informations, cf. section 3.4.1.3.2.

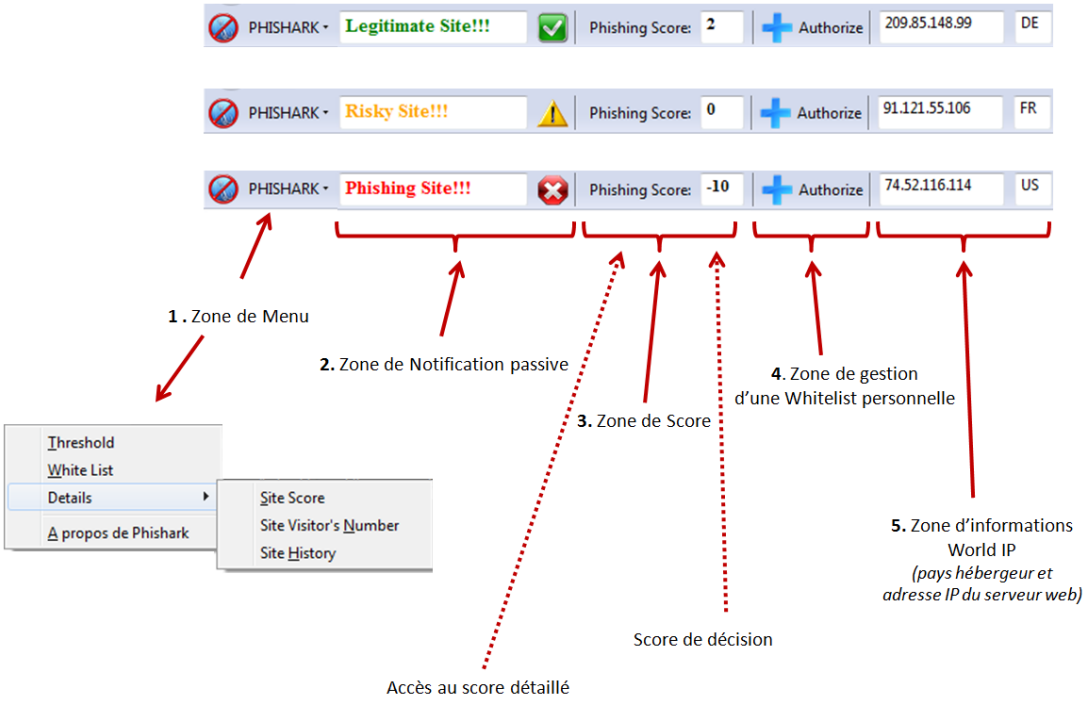


FIGURE 3.4 – Aperçu général de *Phishark*

min./max. peut être définie), obtenir un affichage plus détaillé du score obtenu pour l'URL visitée (cf. figure 3.5), obtenir des informations additionnelles concernant le site visité (p.ex. sa date de mise en ligne) ou encore connaître les informations d'auteur de la barre.
- la zone 2 contient la notification passive de décision de la barre. On aperçoit notamment les changements d'icônes/couleurs selon les 3 décisions possibles : *Légitime*, *Risqué* et *Contrefait*.
- la zone 3 délivre le score de décision. En cliquant sur le bouton *Phishing Score*, une vue plus détaillée est donnée (c.-à-d. on voit apparaître les scores de chaque heuristique/catégorie d'heuristique, tel qu'exposé en figure 3.5). Il s'agit ici d'accéder par un moyen plus rapide aux informations normalement obtenues via le menu de configuration de la barre.
- la zone 4 est un moyen rapide d'ajouter le site visité à la whitelist personnelle de l'Internaute (autrement accessible via le menu de configuration de la barre).
- enfin, la zone 5 indique l'adresse IP du serveur web et le pays hébergeur du site, tels que délivrés par *World IP*.

Enfin, la figure 3.6 montre les messages de notification active délivrés par *Phishark* en cas de sites suspects. Si le site est évalué *Risqué*, un message de type pop-up apparaît pour avertir l'utilisateur. Ce dernier doit alors cliquer sur le bouton *OK* pour continuer sa navigation. Si le site est évalué *Contrefait*, un message pop-up est également affiché pour proposer deux actions à l'utilisateur : soit continuer sa navigation (en cliquant sur *OK*), soit l'arrêter (en cliquant sur *Cancel*). Dans ce dernier cas, un message[1] exposant les dangers du phishing est alors affiché. L'utilisateur peut à nouveau choisir d'ignorer l'avertissement pour continuer sa navigation (bouton *Ignore*), avoir accès à une définition plus détaillée du phishing (bouton *More Information* qui redirige par exemple sur la page `Wikipédia` associée au phishing), ou retourner à sa page d'accueil (bouton *Exit*).

3.5 Phase de vérification et de comparaison aux autres barres d'outils

Après la phase d'étalonnage, nous avons cherché à vérifier l'efficacité de notre moteur de détection. Pour ce faire nous avons testé ses performances sur une whitelist de 500 nouvelles URLs et une blacklist

1. à l'image de celui proposé dans la fonctionnalité anti-phishing de Mozilla Firefox.

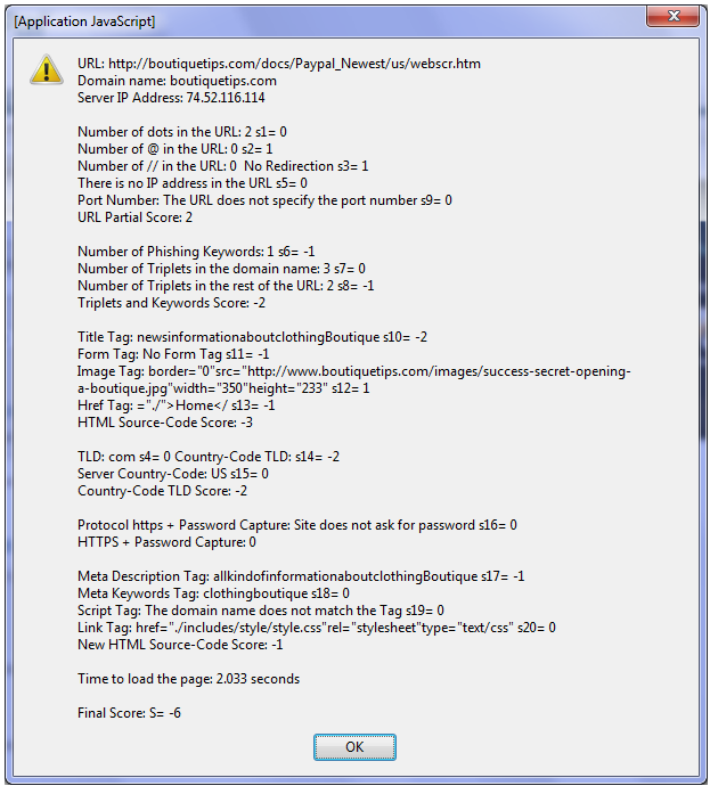
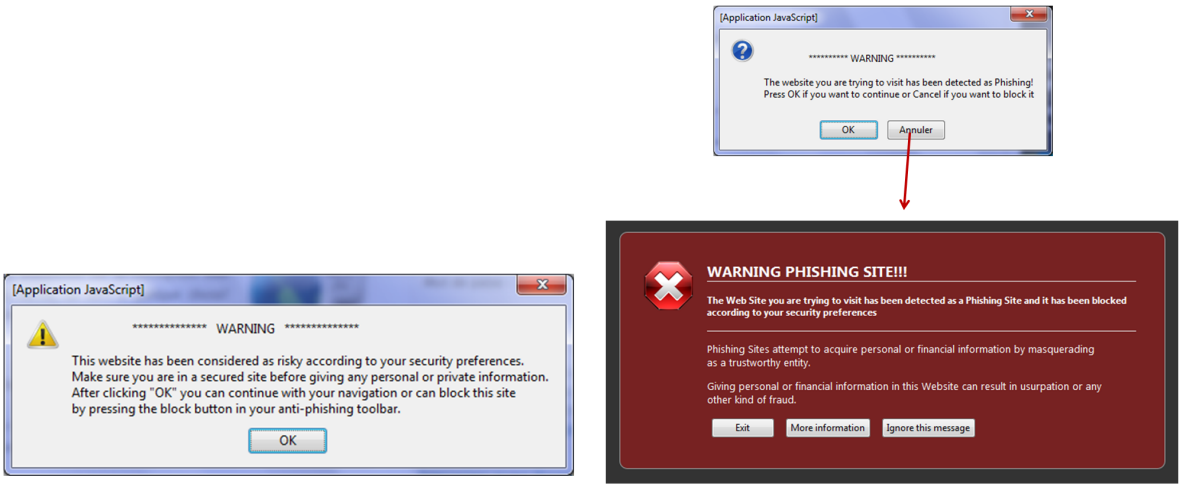


FIGURE 3.5 – *Phishark* : Exemple de scores détaillés des heuristiques



(a) Site Risqué (b) Site Contrefait

FIGURE 3.6 – Messages d'alertes *Phishark* qui requièrent une action de l'utilisateur

de 520 nouvelles URLs. En complément, nous en avons profité pour comparer ses performances aux 4 barres d'outils sélectionnées qui combinent tests heuristiques et blacklist/whitelist (cf. section 3.3).

3.5.1 Performances sur Whitelist

Le tableau 3.6 indique les résultats obtenus lors des tests de performances sur whitelist. En complément, le tableau 3.8 précise l'échelle des scores relevés pour *Phishark*.

Nous constatons que *Netcraft* a réalisé une détection parfaite. Notons toutefois que la barre a certainement été légèrement avantagée, par le fait que le Top 100 des sites les plus visités publié par l'éditeur a contribué à l'élaboration de notre whitelist de test.

Elle est ensuite secondée par *Internet Explorer* qui n'a manqué qu'un seul site : le navigateur nous a en effet fait part d'un message d'erreur, alors qu'au même instant l'ensemble des autres barres réussissaient à accéder au site concerné sans aucun problème.

En troisième position, nous retrouvons *Mozilla Firefox* qui réalise une très bonne détection, malgré 4 sites légitimes indiqués contrefaits à tort. En effet, le navigateur a bloqué ces sites pour cause de certificats problématiques (p.ex. pour un manque de cohérence entre le domaine précisé dans le certificat et le FQDN visité). Précisons que ces problèmes n'ont pas été relevés par *Internet Explorer*.

En quatrième position, nous trouvons *Phishark* qui réalise une bonne performance avec 97.60% de sites détectés légitimes, malgré le non-recours à une liste blanche pour réaliser la détection (à contrario des autres barres mieux classées). Les 12 sites légitimes détectés contrefaits à tort sont majoritairement dus aux heuristiques de la catégorie *Code source HTML*. En effet, sur ces sites légitimes les balises HTML sont incohérentes avec le domaine visité. On peut notamment citer le cas du site Mozilla `http://pv-mirror01.mozilla.org/` - typiquement mentionné dans les listes blanches - qui permet de télécharger des versions logicielles d'outils développés par l'éditeur (p.ex. le client de messagerie Thunderbird). Ce site obtient le score global de -1 pour cause de contenus de balises non soignés (p.ex. la balise <TITLE> contient `indexof`).

On trouve également quelques rares cas qui amènent à des scores négatifs importants, à l'image du site de login Hotmail `https://login.live.com/login.srf?wa=wsignin1.0&rpsnv=11&ct=1314453909&rver=6.1.6206.0&wp=MBI&wreply=http://mail.live.com/default.aspx?rru=inbox&lc=1036&id=64855&mkt=fr-FR&cbcxt=mai&snsc=1` qui obtient un score de -7. Celui-ci cumule en effet un ensemble de comportements suspects : une URL complexe qui utilise de nombreux points, l'utilisation de triplets et mots-clés dits de phishing, l'utilisation de plusieurs TLDs, ou encore des balises sans rapport avec le FQDN (p.ex. la balise <TITLE> contient : `connexion`). De plus, ce site est hébergé sur un nom de domaine global et commun à l'ensemble des services Windows Live de Microsoft, alors que les quelques balises qui pourraient rappeler le FQDN portent uniquement mention du service accédé (p.ex. la balise <META DESCRIPTION> indique `Hotmail nouvelle generation est arrive...`, alors que le FQDN mentionne exclusivement la famille de service Live).

Notons toutefois que le score moyen relevé sur Whitelist pour *Phishark* est de 5.92, avec un écart-type de 3.82. L'intervalle de confiance à 95% (c.-à-d. l'incertitude d'estimation) associé oscille quant à lui entre 5.58 et 6.25 (cf. tableau 3.8). Ces résultats laissent donc présager d'une classification globalement correcte des sites légitimes sur un plus grand nombre d'URLs.

Enfin, en dernière position, nous trouvons *WOT* qui obtient un niveau de détection relativement satisfaisant à 96.20%. Nous notons toutefois que la barre émet une alerte erronée pour 9 sites et, plus surprenant encore, qu'elle ne délivre aucune information pour 10 autres sites.

3.5.2 Performances sur Blacklist

Le tableau 3.7 indique les résultats obtenus lors des tests de performances sur blacklist. En complément, le tableau 3.8 précise l'échelle des scores relevés pour *Phishark*.

En première position, nous trouvons *Phishark* qui réalise la meilleure performance - en comparaison avec les 4 autres barres testées -, en détectant 97.12% des sites de phishing. Ceci s'explique principalement par les nombreux tests heuristiques utilisés par le moteur de détection, dont l'usage semble

Tableau 3.6 – Résultats des tests de performance des 5 barres d'outils sur une Whitelist de 500 URLs

Classement	Barre anti-phishing	Détections positive ou neutre		Faux-positif		Aucune information[1]	
		Quantité	%	Quantité	%	Quantité	%
1	Netcraft	500	100	-	-	-	-
2	Internet Explorer	499	99.80	-	-	1	0.20
3	Mozilla Firefox	496	99.20	4	0.80	-	-
4	**Phishark**	488	97.60	12	2.40	-	-
5	Web of Trust	481	96.20	9	1.80	10	2.00

[1] site déclaré indisponible ou absence de décision de la barre.

moindre dans les autres barres (au profit des blacklists).

Néanmoins, *Phishark* délivre une décision erronée pour 15 sites. Après investigation sur ces mauvaises détections, il apparaît que les URLs concernées sont principalement celles qui utilisent une URL relativement simple quasi-exclusivement constituée d'un FQDN ou d'une arborescence standard de type `http://www.mondomaine.com/dossier1/dossier2/fichier.html`. On peut par exemple citer les cas des sites `http://olegya.fr/` qui obtient un score de 3, `http://www.worldofwarcrajft-account.com/` qui obtient un score de 2, ou encore `http://www.linsenmeier-naturwein.de/naturweine/irs/irsb/SearchTAXERR.php` qui obtient un score de 3. Pour ces cas, nous relevons que les deux premières catégories d'heuristiques (c.-à-d. *Points et caractères spéciaux* et *Triplets et mots-clés dits de phishing*) sont les principales causes de ces détections erronées. En effet, elles sont insuffisamment compensées par les mauvais scores obtenus sur les heuristiques des catégories *Code source HTML* et *Page de login* qui détectent respectivement les balises incohérentes avec le FQDN visité (c.-à-d. toujours liées au FQDN légitime usurpé) et la présence de zones de login dans un site qui ne propose pas de connexion sécurisée.

Malgré ces mauvaises détections, nous notons que le score moyen relevé sur Blacklist pour *Phishark* est de -5.93, avec un écart-type de 3.37, l'intervalle de confiance à 95% associé oscillant entre -6.22 et -5.64 (cf. tableau 3.8). A nouveau, ces résultats rendent assez confiants sur une classification correcte des sites de phishing sur un plus grand nombre d'URLs. Ceci sous réserve bien sûr de l'évolution des techniques de phishing.

Concernant les autres barres d'outils, nous avons remarqué que leurs performances étaient fortement impactées à la baisse dès lors que nous tentions de tester les URLs de phishing dès leur récupération. Nous avons donc réalisé les tests de performances par mini-blocs de 50 à 100 URLs maximum (cf. section 3.3), tant pour minimiser le taux de pertes des URLs (dus à la courte durée de vie des sites de phishing), que pour laisser un délai de synchronisation/mise à jour raisonnable aux blacklists utilisées par ces barres. Ainsi, en moyenne, chaque URL n'était testée que 2 à 3 heures après sa récupération (c.-à-d. le temps nécessaire à la récupération des mini-blocs d'URLs de phishing), en débutant bien évidemment par les premières URLs récupérées.

Nous notons alors que *Netcraft* délivre le deuxième meilleur taux de détection des sites de phishing, avec 90.38% d'alertes. Elle est ensuite suivie par *Mozilla Firefox* qui retourne un taux de détection de 87.31%. En quatrième position arrive *Internet Explorer* avec un taux de 78.08% de détections correctes. Enfin, *WOT* ferme à nouveau les rangs avec une détection de 77.12% des sites contrefaits.

Nous remarquons également que *Mozilla Firefox* et *Internet Explorer* semblent avoir recours à des blacklists différentes. En effet, pour bon nombre de sites de phishing, nous avons vu que certains d'entre eux étaient détectés par une barre et non par l'autre, puis inversement. Néanmoins, au vu des résultats obtenus, il semblerait que celle de *Mozilla Firefox* soit plus performante et/ou plus rapidement mise à jour, de même que le navigateur semble plus rapide pour le chargement des sites.

Aucune barre n'est épargnée par les taux de faux-négatifs qui sont les plus élevés pour *Mozilla Firefox* et *Internet Explorer*, alors respectivement de 12.31 et 18.08%. A contrario, *WOT* délivre moins de faux-négatifs (c.-à-d. 5.19%), mais ceux-ci sont remplacés par une absence totale d'information pour 17.69% des sites testés.

Pour conclure cette phase, précisons que le temps demandé par *Phishark* pour l'évaluation d'un site est en moyenne de l'ordre de 2 à 3 secondes. Ce qui permet d'alerter l'utilisateur dans un délai

Tableau 3.7 – Résultats des tests de performance des 5 barres d'outils sur une Blacklist de 520 URLs

Classement	Barre anti-phishing	Détection négative ou neutre		Faux-négatif		Aucune information[1]	
		Quantité	%	Quantité	%	Quantité	%
1	**Phishark**	505	97.12	15	2.88	-	-
2	Netcraft	470	90.38	44	8.46	6	1.15
3	Mozilla Firefox	454	87.31	64	12.31	2	0.38
4	Internet Explorer	406	78.08	94	18.08	20	3.85
5	Web of Trust	401	77.12	27	5.19	92	17.69

[1] site déclaré indisponible ou absence de décision de la barre.

Tableau 3.8 – Échelle des scores finaux *Phishark* sur Whitelist de 500 URLs et Blacklist 520 URLs

	Échelle des scores relevés (min $\leq$ moyenne $\leq$ max)	Écart-Type	Intervalle de confiance à 95%
WHITELIST	-7 $\leq$ **5.92** $\leq$ 16	3.82	[5.58 ; 6.25]
BLACKLIST	-15 $\leq$ **-5.93** $\leq$ 6	3.37	[-6.22 ; -5.64]

raisonnable (c.-à-d. avant qu'il n'ait pu saisir ses données confidentielles) en cas de site suspect.

3.6 Phase d'identification des heuristiques déterminants

Après avoir apprécié les performances de *Phishark*, comparables à celles des autres barres d'outils, nous avons utilisé notre barre de détection afin d'identifier les heuristiques prédominants pour la décision de légitimité/contrefaçon d'un site.

Pour ce faire, nous avons appliqué notre moteur de détection à 230 URLs de whitelist (extraites des 500 utilisées lors de la vérification) et 230 nouvelles URLs de blacklist. En effet, de par la durée de vie très courte des sites de phishing et le temps nécessité pour tester les performances des 5 barres, il ne nous a pas été possible de réutiliser les URLs précédentes de blacklist ou de procéder à ces tests au cours de la phase de vérification.

Ainsi dans cette nouvelle phase, pour chaque URL testée, nous avons noté les scores intermédiaires de chacune des 6 catégories d'heuristiques, en complément du score global. Les échelles de scores intermédiaires relevés sont détaillées dans les tableaux 3.9 et 3.10, tandis que les échelles de scores finaux sont détaillées dans le tableau 3.11.

Puis, nous en avons déduit les heuristiques prédominants sur whitelist et blacklist, selon la technique énoncée en section 3.3. Les résultats obtenus sont présentés en figure 3.7.

3.6.1 Heuristiques prédominants pour la Whitelist

D'après la figure 3.7, nous constatons que pour les sites légitimes, les catégories *Points et caractères spéciaux* (à hauteur de 95.65%[1]) et *Triplets et mots-clés dits de phishing* (à hauteur de 71.30%[1]) qui portent sur l'étude de l'URL, ainsi que les catégories *Code source HTML* (à hauteur de 66.09%[1]) et *Autres balises HTML* (à hauteur de 63.04%[1]) qui s'intéressent à l'analyse du code source de la page web, sont déterminantes pour l'identification d'un site légitime.

A contrario, les catégories *TLD* et *Page de login* ne semblent pas ici déterminantes, de par leurs taux relevés respectivement à 11.74%[1] et 25.65%[1].

Nous pouvons alors en déduire que les facteurs influençant pour une décision de légitimité sont donc la simplicité de l'URL, et un remplissage adéquat des balises HTML via l'inclusion d'une référence au FQDN visité. Les échelles des scores relevés pour les catégories d'heuristiques associées y sont en effet les plus élevées (cf. tableau 3.9).

1. c-à-d. ces catégories d'heuristiques délivrent un score positif pour le pourcentage d'URL testées mentionné.

Tableau 3.9 – Échelles des scores intermédiaires *Phishark* sur Whitelist de 230 URLs

	Échelle des scores relevés (min ≤ moyenne ≤ max)	Écart-Type	Intervalle de confiance à 95%
Points et caractères spéciaux	-2 ≤ **1.80** ≤ 3	0.77	[1.71 ; 1.90]
Triplets et mots-clés (dits de phishing)	-4 ≤ **0.86** ≤ 4	1.53	[0.66 ; 1.06]
TLD	-2 ≤ **0.05** ≤ 3	0.55	[-0.02 ; 0.12]
Code source HTML	-5 ≤ **0.95** ≤ 5	2.93	[0.57 ; 1.33]
Page de login	-2 ≤ **0.33** ≤ 3	1.59	[0.13 ; 0.54]
Autres balises HTML	-1 ≤ **1.11** ≤ 4	1.16	[0.96 ; 1.26]

Tableau 3.10 – Échelles des scores intermédiaires *Phishark* sur Blacklist de 230 URLs

	Échelle des scores relevés (min ≤ moyenne ≤ max)	Écart-Type	Intervalle de confiance à 95%
Points et caractères spéciaux	-3 ≤ **1.13** ≤ 3	1.20	[0.97 ; 1.28]
Triplets et mots-clés (dits de phishing)	-5 ≤ **-1.16** ≤ 4	2.01	[-1.42 ; -0.90]
TLD	-2 ≤ **-0.26** ≤ 1	0.64	[-0.34 ; -0.17]
Code source HTML	-5 ≤ **-4.60** ≤ 3	1.46	[-4.79 ; -4.41]
Page de login	-2 ≤ **-1.57** ≤ 0	0.83	[-1.67 ; -1.46]
Autres balises HTML	-1 ≤ **-0.42** ≤ 2	0.61	[-0.50 ; -0.34]

A contrario, la notion de connexion sécurisée ne semble pas ici déterminante, ce qui peut aisément s'expliquer par la grande diversité des URLs testées (cf. section 3.3). Appliquée à des URLs de login exclusivement, l'échelle des scores relevés s'en trouverait certainement rehaussée. Il en va de même pour la vérification des informations de localisation du serveur web (via l'analyse du TLD et du pays hébergeur du site). Là aussi, cette faible importance peut s'expliquer par la forte proportion d'URLs utilisant des TLDs génériques[1] et/ou le grand nombre de sites webs gérés par des multinationales. En effet, ces cas de figures rendent difficile une quelconque bonification engendrée par la cohérence des informations TLD / pays hébergeur.

3.6.2 Heuristiques prédominants pour la Blacklist

Pour les sites de phishing, nous constatons que les catégories *Code source HTML* (à hauteur de 96.52%[2]) et *Page de login* (à hauteur de 78.26%[2]) qui s'intéressent à l'analyse du code source de la page web, ainsi que la catégorie *Triplets et mots-clés dits de phishing* (à hauteur de 63.48%[2]) qui porte sur l'étude de l'URL, sont déterminantes pour l'identification d'un site contrefait (cf. figure 3.7).

Une autre catégorie semble d'importance moindre, quoique non négligeable : *Autres balises HTML* avec un taux de 46.96%[2] (cf. figure 3.7).

Enfin, les 2 autres catégories *Points et caractères spéciaux* et *TLD* ne semblent pas ici déterminantes, par leurs taux relevés respectivement à 15.22%[2] et 24.35%[2] (cf. figure 3.7).

Nous pouvons alors en déduire que les facteurs influençant pour une décision de contrefaçon sont donc les incohérences présentes dans les balises HTML (principalement les liens <A HREF> et <IMG>, le titre, ou le contenu des formulaires), l'absence de connexion sécurisée pour les pages de login (toutes les pages de phishing présentent des zones de login), ou la présence des triplets/mots-clés caractéristiques de phishing dans l'URL visitée. Les échelles des scores relevés pour les heuristiques associés indiquent en effet qu'ils présentent les valeurs les plus basses (cf. tableau 3.10).

A contrario des sites légitimes, la complexité des URLs ne semble pas ici déterminante. On peut donc en déduire que cet heuristique n'est indispensable qu'à l'identification des sites légitimes. Par ailleurs, les incohérences entre TLD et pays hébergeur du site ne semblent pas non plus décisives. En effet - de même que vu pour les sites légitimes -, la forte proportion d'URLs utilisant des TLDs génériques explique à nouveau le manque d'efficacité de la catégorie d'heuristiques associée. Notons toutefois que cette catégorie permet tout de même d'aboutir à des scores majoritairement négatifs, ce qui n'était pas le cas pour les sites légitimes.

1. cf. section 4.1.1.1 pour plus d'informations sur ces TLDs.
2. c-à-d. ces catégories d'heuristiques délivrent un score négatif pour le pourcentage d'URL testées mentionné.

Tableau 3.11 – Taux de détections et échelles des scores finaux *Phishark* sur Whitelist et Blacklist de 230 URLs

	Taux de détection correcte ou neutre		Taux de Faux-négatif ou Faux-positif		Échelle des scores relevés	Écart-Type	Intervalle de confiance à
	Quantité	%	Quantité	%	(min ≤ moyenne ≤ max)		95%
WHITELIST	218	94.78	12	5.24	-7 ≤ **5.12** ≤ 16	4.20	[4.57 ; 5.66]
BLACKLIST	227	98.70	3	1.30	-14 ≤ **-6.89** ≤ 3	3.35	[-7.32 ; -6.46]

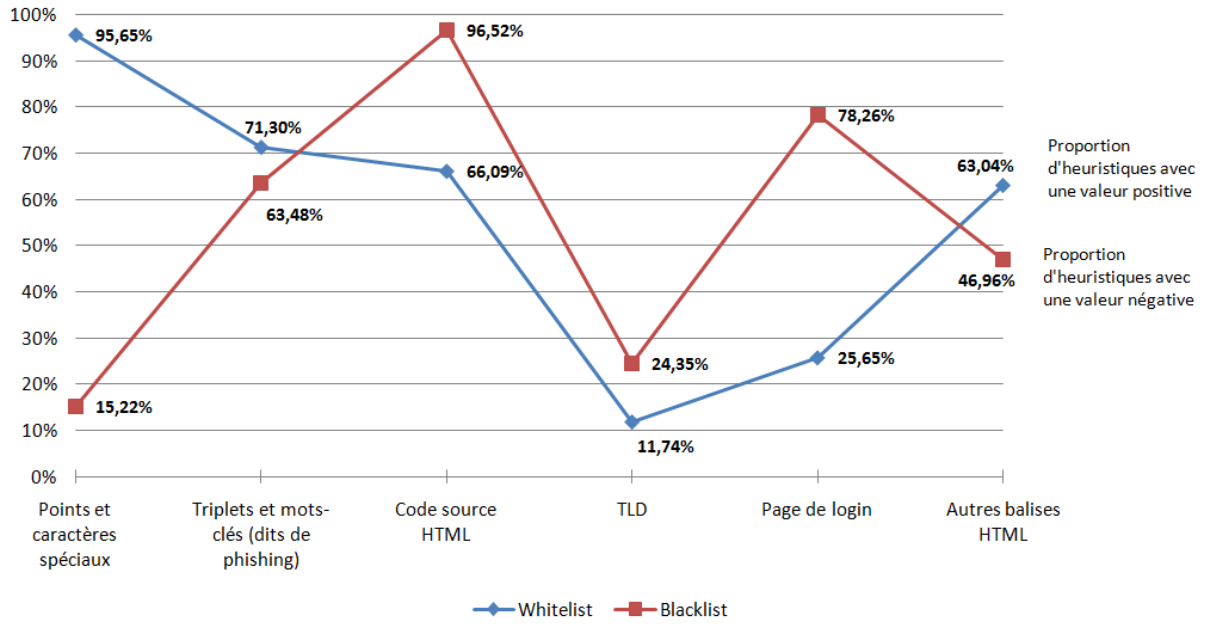


Figure 3.7 – Heuristiques déterminants pour la différenciation des sites légitimes et contrefaits

Enfin, pour conclure notre analyse des heuristiques déterminants, nous pouvons remarquer que les taux de détections obtenus dans cette phase sont d'envergure comparable à ceux relevés lors de la phase de vérification : 94.78% (contre 97.60% précédemment) pour la whitelist, et 98.70% (contre 97.12% précédemment). On note toutefois une légère amélioration pour la détection des sites de phishing et une légère diminution pour la détection des sites légitimes. Cette dernière peut être éventuellement compensée - à l'usage - par la création d'une whitelist personnelle, telle que *Phishark* le propose.

Par ailleurs, le but ultime de cette phase d'analyse est d'aider à affiner les heuristiques et seuils de détection associés, lors d'une étude ultérieure (cf. propositions effectuées en Chapitre 7).

3.7 Discussion sur la pérennité des heuristiques

Au-delà des tests de performances réalisés dans ce chapitre, se pose la question de la pérennité des tests heuristiques utilisés. Parmi les 20 heuristiques utilisés, nous pouvons aisément imaginer qu'il peut être assez facile pour un attaquant de modifier son site contrefait afin qu'il réponde aux exigences de 10 critères (c.-à-d. les heuristiques N° 1, 2, 3, 4, 8, 9, 10, 12, 17 et 18 - cf. tableau 3.5). En effet, si celui-ci a connaissance des tests effectués, il peut modifier son URL afin qu'elle réponde à un schéma plus standard/classique (c.-à-d. peu de points, pas de (@), pas de (//), pas d'adresse IP en guise de FQDN, pas de mots-clés et pas de TLDs "sensibles"), ainsi que le code source de sa page web (c.-à-d. des balises <TITLE>, <META KEYWORDS> et <META DESCRIPTION> cohérentes avec son FQDN), sans que cela lui soit trop coûteux.

Sous réserve que notre moteur de détection soit amélioré afin d'analyser l'ensemble des balises HTML d'un même type (cf. Chapitre 7), il peut ensuite apparaître plus contraignant/problématique de répondre aux exigences de 2 autres critères (c.-à-d. les heuristiques N° 19 et 20 - cf. tableau 3.5). En

effet, il est plus astreignant d'être cohérent au niveau des balises <SCRIPT> et <LINK>, puisque les premières nécessitent un hébergement de l'ensemble des scripts sur le serveur web malveillant, tandis que les secondes nécessitent un hébergement local de la feuille de style par exemple.

Enfin, il est nettement plus compliqué car trop contraignant de répondre aux exigences des 8 critères restants (c.-à-d. les heuristiques N° 5, 6, 7, 11, 13, 14, 15 et 16 - cf. tableau 3.5), voire même quasi-impossible pour certains. En effet, de par les techniques rapides, peu coûteuses et plus transparentes qui sont actuellement utilisées par les attaquants pour confectionner leurs sites contrefaits (c.-à-d. via l'aspiration de sites webs et le maintien d'un maximum de redirections de liens vers le site légitime), l'effort demandé serait considérable. Au-delà de la nécessité de dupliquer l'intégralité du site web usurpé (pour les formulaires, tous les liens <IMG>, <A HREF>), il faudrait également avoir connaissance des triplets de phishing. De plus, les attaquants ont fortement recours aux réseaux de botnets pour héberger leurs sites contrefaits (cf. section 2.3), ceci afin d'être moins facilement identifiables. Cela implique donc souvent l'utilisation d'une redirection de port et/ou une difficulté à être hébergé dans le pays associé au TLD utilisé dans l'URL de phishing. Pour terminer, le critère le plus compliqué à atteindre est certainement celui qui concerne l'utilisation d'une connexion HTTPS. En effet, celle-ci nécessite l'utilisation d'un certificat côté serveur. Si celui-ci est invalide ou incohérent avec le FQDN visité, le navigateur web du client émet aussitôt un message d'alerte (indépendamment d'une quelconque détection du phishing). A contrario, si l'attaquant cherche à détenir un certificat valide, il en devient plus facilement identifiable.

3.8 Problèmes rencontrés

L'une des premières difficultés rencontrées dans cette étude a déjà été en partie exposée en sections 3.3 et 3.5.2. En effet, la durée de vie très courte des sites de phishing introduit un degré de péremption des blacklist élevé. Cet état de fait est à la fois rassurant sur le niveau de réactivité des FAI/hébergeurs qui suspendent les sites émetteurs, mais également fortement contraignant si l'on souhaite utiliser ces blacklists à des fins d'analyses. Par conséquent, le délai entre la récupération des URLs et leurs tests doit être aussi court que possible. De plus, bien que les blacklists proposées par Phishtank et l'APWG soient très régulièrement rafraîchies - p.ex. toutes les 5 minutes sur le site de l'APWG - afin d'informer des derniers sites de phishing détectés, elles n'en comportent pas moins une forte proportion d'URLs périmées. Ainsi, il faut donc effectuer un premier écumage à la récupération des URLs de phishing, pourtant déclarées valides et encore en ligne. En moyenne, le temps de récupération des mini-blocs de 50 à 100 URLs nécessite donc facilement 2 à 3 heures, ce qui n'empêche aucunement que les premières URLs récupérées soient périmées durant ce laps de temps.

Les autres problèmes rencontrés les plus notables et/ou les précautions qu'il a fallu prendre ont principalement concerné la mise en œuvre des heuristiques au sein du moteur de détection. En effet, en premier lieu il faut s'assurer que la page web récupérée est bien valide. En effet, en cas d'indisponibilité de site, il ne sert à rien de lancer la détection de phishing, tant pour ne pas délivrer de décision erronée que ne pas fatiguer l'utilisateur avec de fausses alarmes intempestives. Pour ce faire, nous analysons l'en-tête HTTP récupéré avec la page afin de s'assurer que le code de récupération associé est bien "200" - ce qui signifie que la page a été récupérée avec succès[1] -, avant toute exécution du moteur de détection. Néanmoins, cette technique n'est pas infaillible puisqu'il nous est arrivé de rencontrer des pages d'erreur (p.ex. liées à une indisponibilité temporaire de site) improprement codées au niveau de l'en-tête HTTP (c.-à-d. avec un code d'erreur "200").

Les heuristiques associés aux balises HTML posent également parfois problèmes, puisque les spécificités de certaines langues introduisent des incohérences entre contenu de certaines balises et FQDN visité. En effet, par exemple les FQDN ne contiennent un caractère accentué. Ainsi, à titre d'illustration, une banale vérification de cohérence entre la balise <TITLE> (qui contient `Banque et Assurances - Société Générale`) et son FQDN associé `https://particuliers.societegenerale.fr/index.html` pour le site légitime de login de la Société Générale, nous retournerait un score négatif, à tort. Pour pallier ce problème, nous avons été amenés à implémenter des fonctions spécifiques qui visent à remplacer tous les caractères accentués[2] des balises par leurs pendants sans accent (p.ex. é est remplacé

1. cf. section 6.1.5.2 pour plus de détails sur l'en-tête HTTP et ses codes d'erreurs.
2. aujourd'hui notre implémentation porte sur les caractères "a", "e", "i", "o", "u", "y", "æ", "c" et "n".

par `e`, `ñ` est remplacé par `n`, etc.). Néanmoins, la liste utilisée aujourd'hui n'est pas exhaustive. Elle se doit donc d'être enrichie pour éviter ce travers.

Enfin, un des soucis rencontrés pour lequel nous n'avons pas de solution actuellement concerne la redirection de sites webs. En effet, certaines URLs légitimes procèdent à une redirection automatique vers d'autres URLs. C'est le cas par exemple du site de login Hotmail, où `www.hotmail.com` n'est qu'un alias qui redirige vers `https://login.live.com/login.srf?wa=wsignin1.0&rpsnv=11&ct=1314530844&rver=6.1.6206.0&wp=MBI&wreply=http:%2F%2Fmail.live.com%2Fdefault.aspx&lc=1036&id=64855&mkt=fr-FR&cbcxt=mai&snsc=1`. Ainsi l'analyse est effectuée uniquement à partir de cette dernière URL et de la page web associée. Tels qu'exposés précédemment (cf. section 3.4), la complexité et le manque de lien entre contenu de la page et FQDN de l'URL de redirection entraîne alors un score négatif pour ce site (de -7). Ceci n'aurait pas été le cas, si nous avions utilisé l'URL originale saisie par l'utilisateur (c.-à-d. `www.hotmail.com`). Néanmoins, parce que des sites de phishing pourraient également utiliser cette technique en contournement (c.-à-d. en présentant systématiquement une URL simple avant une redirection vers une URL plus complexe), nous ne pensons pas qu'il soit vraiment souhaitable de chercher à remédier à ce problème.

3.9 Synthèse du chapitre

Dans ce chapitre, nous avons évalué l'efficacité et la prédominance des tests heuristiques utilisés au sein de barres anti-phishing, pour la détection des sites contrefaits visités par l'Internaute. Pour ce faire nous avons implémenté notre propre barre anti-phishing, dont le moteur de détection est exclusivement basé sur 20 tests heuristiques qui analysent aussi bien l'URL que le contenu du code source HTML de la page web.

Nous avons ainsi prouvé que l'utilisation des heuristiques - qu'ils se rapportent à l'étude de l'URL ou l'analyse du code source HTML -, est primordiale pour une détection efficace des sites de phishing dès leur apparition. Toutefois, il apparaît clairement que tous ces heuristiques ne sont pas égaux dans la détection. En particulier, la simplicité des URLs légitimes et l'analyse du code source HTML associé sont essentielles pour une décision de légitimité. Cette même analyse du code source HTML et la vérification d'une connexion sécurisée sur une page de login sont à leur tour vitales pour une décision de contrefaçon. Les résultats que nous obtenons ici nous permettent également de distinguer plusieurs pistes d'amélioration des tests heuristiques utilisés. Celles-ci sont exposées dans le Chapitre 7.

Pour terminer, nous mesurons également la principale faiblesse des heuristiques, à savoir : leurs taux de FPR/FNR. Cette faiblesse peut à elle seule aisément expliquer l'utilisation de whitelist/blacklist additionnelles, dans le but d'obtenir une détection moins encline aux erreurs. Toutefois, l'un des inconvénients majeurs introduits par ces listes blanches/noires est l'opportunité supplémentaire qu'elles offrent pour la corruption du moteur de détection.

Dans la même veine, quelque soit l'efficacité de leur détection, les barres anti-phishing n'en demeurent pas moins assujetties aux faiblesses du système sur lequel elles reposent, à savoir : le navigateur web, le poste client et le réseau auquel il appartient. La Seconde Partie de ce mémoire essaie de se pencher sur une fraction de cet épineux problème.