Daniel Mider
Jan Garlicki
Wojciech Mincewicz

Pozyskiwanie informacji z Internetu metodą Google Hacking – biały, szary czy czarny wywiad?

Wyszukiwarka Google jest równie powszechnie (i niemal bezalternatywnie) używana, co nieznana. Nie docenia się jej możliwości w zakresie pozyskiwania tzw. danych wrażliwych osób i instytucji. Zapytanie skierowane do Google, odpowiednio sformułowane, pozwala na wyszukiwanie stron usuniętych i archiwalnych, odtwarzanie struktury witryny internetowej lub struktury sieci wewnętrznej, dostęp do parametrów konfiguracyjnych serwerów, pozyskiwanie informacji celowo zabezpieczonych przed dostępem do nich (ang. *paywall* – 'hasła'), pozyskanie nazw użytkowników i haseł, ich numerów identyfikacyjnych (np. numerów kart płatniczych, numerów PESEL) oraz dostęp do parametrów konfiguracyjnych urządzeń (serwerów, kamer, routerów i innych) w celu przejęcia nad nimi kontroli. Te działania określa się w literaturze przedmiotu jako Google Hacking (GH), Google Dorks (GD) lub (rzadziej) – Google Scanning (GS) bądź Engine Hacking (EH). Pod tymi terminami rozumie się takie formułowanie zapytań dla przeglądarki Google, aby udostępniała dane, do których użytkownik jest nieuprawniony – w sensie etycznym, prawnym bądź w obu tych sensach[1].

Pojęcie Google Hacking zostało wprowadzone przez autorytet w tej dziedzinie – Johnny'ego Longa[2] – i oznacza biegły sposób korzystania z wyszukiwarki Google. Pojęcie Google Dork natomiast oznacza osobę, którą charakteryzuje nieudolność w zakresie zabezpieczania treści zamieszczanych online, głównie witryn internetowych. Te słabości mogą zostać łatwo ujawnione przez Google (w oryg. *An inept or foolish person as revealed by Google*). Jak wskazuje autor, centrum znaczeniowe leksemu „dork" uległo z czasem przesunięciu i obecnie oznacza osobę, która za pomocą Google pozyskuje informacje poufne[3].

---

[1] Por.: J. Long, *Google Hacking for Penetration Testers*, Rockland 2007, s. 534. Podobną definicję zawiera również słownik Collinsa: https://www.collinsdictionary.com/submission/9695/google+dorks [dostęp: 26 I 2018]. Równolegle oprócz takich definicji jak ta pojawia się wiele definicji wykorzystujących socjolekt informatyczny, jednak z grubsza oznaczających to samo, na przykład rozumienie GH jako „logiczne exploity wyszukiwarkowe" lub „pozyskiwanie swoistego wirtualnego notatnika" (ang. *virtual notebook*). Zob. *Google Hacking – w służbie hakerom*, „Haker.edu.pl", 10 VII 2015, w: https://haker.edu.pl/2015/07/10/google-hacking-google-dorks/ [dostęp: 26 I 2018]. W tekście używa się wymiennie dwóch najbardziej rozpowszechnionych pojęć – Google Hacking i Google Dorks. Patrz: *Roll Call Release. Intelligence for Police, Fire, EMS, and Security Personnel*, 7 VII 2014, https://info.publicintelligence.net/DHS-FBI-NCTC-Google Dorking.pdf [dostęp: 26 I 2018].

[2] Jest to znany twórca nieistniejącej już strony http://johnny.ihackstuff.com; obecnie treści zawarte na tej stronie znalazły się pod adresem http://www.hackersforcharity.org/ghdb/, tj. Google Hacking Database. Znany również pod pseudonimami „j0hnny" i „j0hnnyhax".

[3] J. Long, *The Google Hacker's Guide. Understanding and Defending Against the Google*

Historia GH/GD rozpoczęła się wraz ze zdefiniowaniem nazwy zjawiska w grudniu 2002 r. przez pioniera tej metody Johnny'ego Longa, choć jako technika wykorzystywana przez hakerów istniała już od roku 2000[4]. W 2004 r. powstało pierwsze narzędzie informatyczne służące samoobronie przed GH – SiteDigger (w wersji 1.0), a rok później udostępniono programy Google Hack Honeypot i MSNPawn (właśc. Search.msn). W 2005 r. zostało po raz pierwszy wydane fundamentalne, wielokrotnie wznawiane dzieło *Google Hacking for Penetration Testers*[5]. Pojęcie GH/GD może być nieco mylące. Analogiczne informacje, jak wymienione, można pozyskiwać również za pomocą innych wyszukiwarek, jak na przykład Bing, Yahoo, Yandex i DuckDuckGo[6]. Różnice w strukturze samych zapytań (tj. komend konstruowanych przez użytkownika z wykorzystaniem predefiniowanych elementów nazywanych operatorami) są zazwyczaj niewielkie i dotyczą ich nazw, sposobu zapisu niektórych z nich (operatory George'a Boole'a), a także niektórych operatorów obecnych wyłącznie w danej wyszukiwarce[7].

W literaturze przedmiotu występują różne propozycje klasyfikacji i systematyzacji rodzajów zapytań GH. Na przykład J. Long wymienia aż czternaście typów GH, podczas gdy Flavio Toffalini wraz z zespołem proponuje cztery następujące, ogólne kategorie zapytań: 1 – lokalizujące serwery (wersje oprogramowania), 2 – lokalizujące katalogi wrażliwe, 3 – lokalizujące pliki, które zawierają hasła oraz 4 – lokalizujące pliki (logi) z błędami systemowymi[8]. Inny pomysł taksonomizacji opiera się na następującej triadzie odwołującej się do właściwości samych zapytań (nie zaś ich funkcji): 1 – odnoszące się do struktury URL (ang. *Uniform Resource Locator*), 2 – odnoszące się do rozszerzenia pliku, 3 – odnoszące się do zawartości pliku[9]. Spektrum dopuszczalności gromadzenia informacji rozciąga się od dozwolonych, bo legalnych i etycznych, do niedozwolonych, będących nielegalnymi i nieetycznymi sposobami jej pozyskiwania. Przyjęło się wyróżnianie trzech następujących klas pozyskiwania informacji

Hacker, http://pdf.textfiles.com/security/googlehackers.pdf [dostęp: 26 I 2018].

[4] *Smart searching with googleDorking*, „Exposing the Invisible", https: //exposingtheinvisible.org/guides/google-dorking/#dorking-operators-across-google-duckduckgo-yahoo-and-bing [dostęp: 26 I 2018].

[5] J. Long, B. Gardner, J. Brown, *Google Hacking for Penetration Testers*, Amsterdam i in. 2005.

[6] Warto zasygnalizować istnienie wyszukiwarek Shodan, Censys oraz Zoomeye, które są projektami szczególnie interesującymi, nowatorskimi i nietypowymi w kategoriach tego typu wyszukiwarek. Są to wyszukiwarki Internet of Things – komputerów i urządzeń sieciowych (w istocie, z informatyczno-technicznego punktu widzenia, to skanery portów online). Wyszukiwarki znajduje się odpowiednio pod adresami: https://www.shodan.io/; https://censys.io/; https://zoomeye.org.

[7] Na przykład w Yahoo i Bing istnieje operator „language:" używany z kodem określonego języka, pozwalający na znalezienie poszukiwanego terminu właśnie w tym języku. Z kolei w Google funkcjonują operatory „book:" oraz „maps:" pozwalające na wyszukiwanie odpowiednio plików zdefiniowanych jako książki lub mapy. Operator „cache:" istnieje tylko w Google, natomiast operatory „site:", „intitle:" i „filetype:" są uniwersalne. Autorzy dysponują porównawczym zestawieniem operatorów (jest ono stale uzupełniane). W przypadku zaistnienia potrzeby autorzy mogą je udostępnić

[8] F. Toffalini i in., *Google Dorks: Analysis, Creation, and new Defenses*, w: *Detection of Intrusions and Malware, and Vulnerability 2016*, J. Caballero i in. (red. nauk.), San Sebastian 2016, s. 255–275.

[9] Tamże.

określanych mianem białego, szarego i czarnego wywiadu[10]. Biały wywiad to taki sposób gromadzenia danych, który nie budzi wątpliwości zarówno etycznych, jak i prawnych. Wskazuje się, że 80 proc. informacji jest współcześnie pozyskiwanych w drodze eksploracji źródeł otwartych, jawnych, na przykład: państwowych, prasowych i prywatnych. Z kolei przeciwstawny typ źródeł – czarny wywiad – to czynności jednoznacznie nieetyczne i nielegalne. Do tej grupy należy zaliczyć m.in. instalację podsłuchów osób, pomieszczeń, telefonów, włamania, kradzieże tożsamości i parametrów biometrycznych, łamanie zabezpieczeń kryptograficznych oraz pozyskiwanie informacji za pomocą szantażu i korupcji. Jak wskazuje Kazimierz Turaliński, ten typ wywiadu dostarcza 5 proc. wszystkich informacji. Z kolei szary wywiad lokuje się pomiędzy dwoma opisanymi – są to działania, które nie mogą być zaklasyfikowane jednoznacznie. Są one legalne, ale nieetyczne. Obejmują m.in. inwigilację (obserwację i monitoring), infiltrację oraz działania o charakterze socjotechnicznym. Za pomocą szarego wywiadu pozyskuje się około 15 proc. informacji[11]. Zaproponowany podział treści artykułu na części dotyczące techniki białego, szarego i czarnego wywiadu ma walor tyleż problemowy, co porządkujący. Może posłużyć rozróżnieniu tego, co z perspektywy systemów normatywnych (etyki i prawa) jest dozwolone lub nie. Do białego wywiadu zostały zaliczone techniki wyszukiwania informacji w Google, które nie budzą kontrowersji etycznych i prawnych, tj. wyszukiwanie informacji archiwalnych i usuniętych oraz pozyskiwanie ogólnodostępnych danych osobowych. Przez szary wywiad rozumie się natomiast pozyskiwanie wrażliwych danych osobowych, uzyskiwanie dostępu do zabezpieczonych treści, dostęp do urządzeń online i parametrów konfiguracyjnych urządzeń (np. pliki drukarek), a także odtwarzanie struktury strony witryny lub struktury sieci wewnętrznej. Czarny wywiad GH obejmuje pozyskiwanie list użytkowników i haseł, zbiorów z informacjami wysoko wrażliwymi (podstawowych danych osobowych, takich jak numery kart kredytowych, numery PESEL, numery ubezpieczeń) oraz dostęp do urządzeń online, zarówno prywatnych, jak i instytucjonalnych (instytucji państwowych i przedsiębiorstw prywatnych), w tym sieci monitoringu (kamer).

Warto podkreślić, że wykorzystywanie technik i narzędzi analizowanych w niniejszym artykule może rodzić odpowiedzialność karną. Szczególne znaczenie ma pierwszy spośród tak zwanych „paragrafów hakerskich" *Kodeksu karnego*, tj artykuł 267. Wykorzystując GH w sposób nieuprawniony, bez pisemnej zgody i wiedzy podmiotu, wobec którego takie działania są podejmowane, narusza się artykuł 267 kk, nakładający karę pozbawienia wolności do lat dwóch, karę ograniczenia wolności lub grzywnę wskutek pozyskania bez uprawnienia informacji wrażliwej bądź uzyskania dostępu do całości lub części systemu informatycznego. Prawo w tym zakresie zostało jednak nieco zliberalizowane – 11 kwietnia 2017 r. prezydent Andrzej Duda podpisał nowelizację kodeksu karnego. Tekst „paragrafów hakerskich" (od 267 do 269b) uzupełniono

[10] Por. A.W. Dorn, *United Nations Peacekeeping Intelligence*, w: *National Security Intelligence*, L.K. Johnson (red. nauk.), Oxford 2010, s. 280.

[11] K. Turaliński, *Wywiad gospodarczy i polityczny. Podręcznik dla specjalistów ds. bezpieczeństwa, detektywów i doradców gospodarczych*, Warszawa 2015, s. 31–33.

o tak zwane kontratypy. Dotychczas prawo przewidywało bezwzględne orzekanie kar za (…) *wytwarzanie, pozyskiwanie, zbywanie lub udostępnianie innym urządzeń lub programów przystosowanych do popełnienia przestępstw komputerowych* (art. 269b), w tym konfiskatę rozszerzoną. Taka konstrukcja prawna nie uwzględniała działań podejmowanych przez zawodowych testerów zabezpieczeń oraz badaczy opracowujących metody zabezpieczeń. Po nowelizacji działalność testerów bezpieczeństwa sieci jest w pełni legalna[12].

Pracy nad niniejszym artykułem przyświecały dwa cele. Po pierwsze, uświadomienie czytelnikom zagrożeń i możliwości związanych z badanym zjawiskiem, a po drugie – zawarcie w nim informacji dotyczących takich elementów praktycznych, które pozwolą na samodzielne wykorzystanie opisanych technik w celach publicznie pożytecznych. Z tym jednak ważnym zastrzeżeniem, że nie przedstawiono tu wyczerpującego kompendium technik służących pozyskiwaniu danych z sieci Internet, lecz techniki typowe, umożliwiające dalszy samodzielny rozwój w tym zakresie. Wysiłki poznawcze zogniskowano na wyszukiwarce Google, ponieważ dominuje ona w Internecie – blisko dziewięć na dziesięć zapytań (87,16 proc.) jest kierowanych właśnie do niej[13]. Uzupełnienie artykułu stanowi analiza pozyskiwania informacji za pomocą programu FOCA[14] (*Fingerprinting Organizations with Collected Archives*). Roboty Google nie indeksują tzw. metaznaczników dokumentów[15] (zawierających między innymi informacje o datach ich wykonania, modyfikacji czy personaliach osób, które je utworzyłyitp.), program FOCA natomiast umożliwia ich pozyskanie.

---

[12] Paragraf 1a wprowadzony do art. 269b kk reguluje tę kwestię następująco: „Nie popełnia przestępstwa określonego w § 1, kto działa wyłącznie w celu zabezpieczenia systemu informatycznego, systemu teleinformatycznego lub sieci teleinformatycznej przed popełnieniem przestępstwa wymienionego w tym przepisie albo opracowania metody takiego zabezpieczenia". Szczegółową analizę tzw. paragrafów hakerskich przeprowadził Filip Radoniewicz. Zob. F. Radoniewicz, *Odpowiedzialność karna za przestępstwo hackingu,* https://www.iws.org.pl/pliki/files/Filip%20Radoniewicz%2C%20Odpowiedzialno%C5%9B%C4%87%20karna%20za%20przest%C4%99pstwo%20hackingu%20%20121.pdf [dostęp: 5 VI 2018] – stan prawny na rok 2013, przed nowelizacją z kwietnia 2017 r. W 2017 r. Ministerstwo Cyfryzacji przedstawiło Strategię Cyberbezpieczeństwa na lata 2017–2022, w której jest rozpatrywana możliwość prawnego uregulowania systemu bug-bounty. Należy zatem domniemywać, że w naszym kraju powstanie szybka i mało kosztowna metoda testowania zabezpieczeń technologii informacyjnej. Zob. M. Długosz, *Legalny hacking w Polsce. (Analiza)*, 30 V 2017, http://www.cyberdefence24.pl/legalny-hacking-w-polsce-analiza [dostęp: 5 VI 2018].

[13] C. Glijer, *Ranking światowych wyszukiwarek 2017: Google, Bing, Yahoo, Baidu, Yandex, Seznam*, „K2 Search", 25 VII 2017, http://k2search.pl/ranking-swiatowych-wyszukiwarek-google-bing-yahoo-baidu-yandex-seznam/ [dostęp: 26 I 2018].

[14] Program FOCA jest zautomatyzowanym narzędziem analitycznym stworzonym przez firmę ElevenPaths w celu wyszukiwania, pobierania i analizowania dokumentów dla pozyskania informacji cyfrowych pozostawionych przez użytkowników świadomie bądź nieświadomie w zasobach Sieci 1.0 i 2.0. Szczegółowe informacje oraz darmowa wersja są dostępne pod adresem: https://www.elevenpaths.com/labstools/foca/index.html [dostęp: 29 I 2018].

[15] Spośród wyszukiwarek internetowych metaznaczniki dokumentów są indeksowane jedynie przez Bing, które wprowadziło przeznaczny do tego celu operator „meta:".

1. Google Hacking jako biały wywiad

Usługi Google zaklasyfikowane jako biały wywiad są następujące: wyszukiwanie stron usuniętych i archiwalnych, wyszukiwanie informacji o użytkownikach (wyszukiwanie adresów e-mail, wyszukiwanie użytkowników serwisów społecznościowych, wyszukiwanie numerów telefonów – usługa ograniczona wyłącznie do terenu Stanów Zjednoczonych Ameryki) oraz ułatwienia wyszukiwania informacji merytorycznych (wyszukiwanie w hasztagach, wyszukiwanie stron podobnych lub słów związanych, wyszukiwanie definicji pojęć – encyklopedycznych i słownikowych, wyszukiwanie stron zamieszczających odnośniki do interesującej nas strony lub materiału, wyszukiwanie określonych typów plików)[16].

Wyszukiwanie stron usuniętych i archiwalnych

Najbardziej praktyczną, a jednocześnie najciekawszą, usługą dostarczaną przez Google, którą można zakwalifikować do białego wywiadu, jest możliwość wyszukiwania stron usuniętych lub archiwalnych. Można tego dokonać za pomocą operatora „cache:". Zasadą jego działania jest wyświetlanie usuniętej wersji danej witryny przechowywanej przez Google w pamięci podręcznej (ang. *cache*). Typowa składnia jest następująca:

cache:www.inp.uw.edu.pl

Po wpisaniu powyższej frazy do wyszukiwarki Google w tym przypadku pozyska się poprzednią wersję strony Instytutu Nauk Politycznych Uniwersytetu Warszawskiego. Polecenie pozwala na wyświetlenie pełnej wersji (HTML lub zapisanej w innym języku skryptowym, a pojawiającej się „jak jest"), wersji tekstowej oraz źródła skryptu. Podawany jest również dokładny czas (data, godzina, minuta i sekunda), w jakim pająk Google dokonał indeksacji. Strona jest wyświetlana w postaci pliku graficznego, można ją wykorzystać do wyszukiwania (za pomocą skrótu CTRL+F). Wyniki polecenia „cache:" zależą od tego, z jaką częstotliwością robot Google indeksuje strony. Samodzielne ustawienie przez autora znacznika z określoną częstotliwością wizytacji w nagłówku dokumentu HTML jest uznawane przez Google za opcjonalne i na ogół ignorowane na rzecz wielkości uzyskanego współczynnika PageRank, który stanowi główny czynnik częstości indeksowania strony. Jeśli zatem jakaś strona była wymieniana pomiędzy odwiedzinami robota Google, wówczas nie zostanie ona zaindeksowana, a przez to – odczytana poleceniem „cache:". Szczególnie wdzięcznym obiektem dla testowania tej funkcji są blogi, konta serwisów społecznościowych oraz strony portali i wortali internetowych, aktualizowane z dużą częstotliwością. Usunięte informacje zamieszczone przez pomyłkę lub w toku pracy

[16] Podstawowe reguły wyszukiwania w Google, obejmujące między innymi operatory logiczne George'a Boole'a oraz operatory zakresowe, zostaną przytoczone w tekście tam, gdzie to będzie konieczne. Zrezygnowano natomiast z systematycznego wyłożenia podstaw wyszukiwania w Google.

nad stroną czy też wymagające w danym momencie usunięcia, mogą być w łatwy sposób przywołane. Te funkcje znajdują zastosowanie zarówno w praktykach szarego, jak i czarnego wywiadu. Nieostrożność administratora strony czy portalu może narazić go na upublicznienie niechcianych wiadomości[17].

Wyszukiwanie informacji o użytkownikach

Wyszukiwanie informacji o użytkownikach tylko z pewnymi zastrzeżeniami może być uznane za GH. W tym przypadku to określenie wydaje się na wyrost. Służą do tego zaawansowane operatory poprawiające i uszczegóławiające wyniki wyszukiwania[18]. Operator „@" (at) służy do wyszukiwania użytkowników w serwisach społecznościowych – indeksowane są m.in. Twitter, Facebook, Instagram. Przykładowy sposób zastosowania tego operatora jest następujący:

inurl:twitter@mider

W serwisie Twitter będzie wyszukiwany użytkownik „mider".

Przypuśćmy, że znamy miejsce pracy osoby poszukiwanej, np. Instytut Nauk Politycznych UW, i jej nazwisko. Zamiast żmudnie przeszukiwać strony wskazanej instytucji, można wpisać zapytanie, wiedząc, jaka jest konstrukcja adresu poczty elektronicznej, oraz przypuszczając, że w nazwie tego adresu powinno się znajdować co najmniej nazwisko osoby poszukiwanej:

site:inp.uw.edu.pl "*mider@uw.edu.pl"

Można również użyć metody mało wysublimowanej i zadać pytanie o adresy poczty elektronicznej w sposób niżej zaprezentowany, licząc na łut szczęścia oraz brak profesjonalizmu administratora:

emaile.xlsx
filetype:xls +emaile

[17] Wartościowa wydaje się w tym kontekście usługa systematycznego archiwizowania stron internetowych w ramach projektu Internet Wayback Machine, działającego już od ponad półtorej dekady (od 2001 r.). Baza danych znajduje się pod adresem https://web.archive.org i zawiera wiele przeszłych wersji stron www, możliwych do wyszukiwania według licznych, szczegółowych rodzajów zapytań. O ile operator „cache:" indeksuje tylko poprzednią wersję strony, o tyle wymieniony projekt – także wcześniejsze wersje (choć niesystematycznie).

[18] Pominięto techniki oczywiste, zastosowanie zapytań bezpośrednich polegających na wpisywaniu do wyszukiwarki nazwisk, pseudonimów i znanych lub przypuszczalnych nickname'ów oraz identyfikatorów, a także wykorzystanie operatorów modyfikujących zapytania, w tym operatorów logicznych Boole'a. Zaprezentowano metody mniej znane i mające istotne ograniczenia.

Adresy poczty elektronicznej można również spróbować zebrać z danej strony internetowej za pomocą zapytania:

site:inp.uw.edu.pl intext:e-mail

Na stronach INP UW zostanie wówczas wyszukane słowo „e-mail". Wyszukiwanie adresów poczty elektronicznej ma ograniczone zastosowanie i najczęściej wymaga zgromadzenia już określonych informacji o użytkowniku.

Z kolei wyszukiwanie numerów telefonów dzięki operatorowi „phonebook:" Google jest ograniczone wyłącznie do abonentów USA. Przykładowe zapytanie można sformułować następująco:

phonebook:John Doe New York NY

Wyszukiwanie informacji o użytkownikach może się odbywać za pomocą usługi Google „Wyszukiwanie obrazem". Umożliwia ona odnalezienie tożsamych lub podobnych (układ kształtów i barw) fotografii na stronach indeksowanych przez Google.

Wyszukiwanie adresów e-mail jest w Google dość uciążliwe w porównaniu z takimi aplikacjami, jak Maltego czy The Harvester, jednak jest możliwe.

Wyszukiwanie informacji merytorycznych

Google wprowadziło kilka przydatnych ułatwień, między innymi w postaci operatora „related:" powodującego wyświetlenie listy stron „podobnych" do określonej strony internetowej. Przy czym podobieństwo jest oparte na związkach funkcjonalnych, a nie logicznych i merytorycznych:

related:www.inp.uw.edu.pl

W powyższym przypadku są wyświetlane strony innych ośrodków naukowych. Ten operator działa jak przycisk „Podobne strony" w wyszukiwaniu zaawansowanym Google. Analogicznie, zapytanie „info:" umożliwia wyświetlenie informacji o danej stronie www. Są to informacje udostępniane przez autorów danej strony, wprowadzone w nagłówku strony (<head>) w metaznacznikach opisu (<meta name= "Description"…"). Oto przykład zastosowania:

info:inp.uw.edu.pl

Przydatne, szczególnie w pracy naukowej, bywa zapytanie „define:" umożliwiające pozyskanie definicji pojęć z takich źródeł, jak encyklopedie i słowniki online. Sposób wykorzystania tej funkcji jest następujący:

define:political participation

Operatorem o uniwersalnym zastosowaniu jest tylda (~). Dzięki niej można wyszukać słowa pokrewne lub podobne (synonimiczne):

~nauki ~polityczne

Wpisanie powyższego zapytania umożliwi wyszukanie zarówno stron, na których znajdują się wyrazy „nauki" oraz „polityczne", jak i pojęcia synonimicznego – „politologia".

Natomiast operator modyfikujący zapytanie „link:" ogranicza zakres wyszukiwania do odnośników podanych na danej stronie. Dzięki niemu można sprawdzić, czy ktoś zamieszcza odnośniki do interesującej nas strony lub pliku. Na przykład:

link:www.inp.uw.edu.pl

Operator, o którym mowa, działa jednak wadliwie. Nie wyświetla wszystkich wyników i rozszerza kryteria wyszukiwania.

Hasztagi (ang. *hashtag*) są to słowa stanowiące formę znacznika (ang. *tag*) i umożliwiające grupowanie wiadomości poprzedzone znakiem „#". Obecnie są one używane głównie w serwisie Instagram, ale też w takich serwisach, jak Facebook, Google+, Tumblr i Wykop. Google umożliwia wyszukiwanie jednocześnie w wielu serwisach bądź w serwisach wskazanych. Typowe zapytanie dla dowolnych serwisów jest formułowane następująco:

#polityka

Operator „around(n)" umożliwia wyszukanie dwóch słów znajdujących się w określonej odległości od siebie. Na przykład efektem zapytania:

google around(4) hacking

będzie odnalezienie stron, które zawierają te dwa słowa („google" oraz „hacking"), ale są oddzielone od siebie czterema słowami.

Niezwykle przydatne jest wyszukiwanie według typów plików, ponieważ Google indeksuje materiały również ze względu na format, w jakim zostały zapisane. Służy do tego operator „filetype:". Obecnie jest obsługiwany bardzo szeroki zakres plików[19].

Spośród wszystkich wyszukiwarek Google dostarcza najbardziej złożonego zakresu operatorów umożliwiających prowadzenie białego wywiadu. Pozostałe

[19] Aktualna lista typów plików znajduje się pod adresem: https://support.google.com/webmasters/answer/35287?hl=en [dostęp: 26 I 2018].

wyszukiwarki zawierają nieco mniej przydatnych operatorów, w związku z czym nie gwarantują tak precyzyjnego wyszukiwania. Jednak do celów białego wywiadu należy rekomendować takie narzędzia, jak Maltego[20], Oryon OSINT Browser[21] i program FOCA, umożliwiające zautomatyzowane pozyskiwanie danych i niewymagające znajomości operatorów. Mechanizm działania programu jest bardzo prosty: za pomocą odpowiedniego zapytania kierowanego do wyszukiwarki Google, Bing i Exalead odnajduje on dokumenty opublikowane przez interesującą nas instytucję i analizuje metadane zawarte w tych dokumentach. Potencjalnym zasobem informacji dla programu jest każdy plik z dowolnym rozszerzeniem, np. doc, pdf, ppt, odt, xls czy JPG. Jest to usługa dostarczona przez FOCĘ, najbardziej praktyczna i jednocześnie najłatwiejsza do uzyskania, którą można zakwalifikować do białego wywiadu.

Istotne jest zatem, aby przed udostępnieniem pliku zadbać o „sprzątnięcie" metadanych. W poradnikach internetowych bez problemu można odnaleźć co najmniej kilka sposobów na to, jak pozbyć się niechcianych informacji. Nie jest możliwe, aby w sposób aprioryczny wskazać jeden właściwy sposób, ponieważ wszystko jest uzależnione od indywidualnych preferencji użytkowników. Autorzy niniejszej publikacji zalecają, aby przed udostępnieniem dokumentu w pierwszej kolejności zapisać plik w formacie, który nie przechowuje metadanych lub przechowuje je w bardzo ograniczonym zakresie. Jest więc pożądane, aby np. dokument doc.[22] przekonwertować na format txt. lub rtf., obrazy zaś dotychczas zapisywane jako JPG upublicznić w formacie PNG.

Internet oferuje także bogate zasoby darmowych programów „czyszczących" metadane, głównie w przypadku obrazów, gdzie za sprawdzony i pożądany może uchodzić ExifCleaner[23]. W przypadku plików tekstowych jest mile widziane ręczne wykonanie tej czynności. Jej przebieg zależy od pakietu, jakim się dysponuje. Sposobem na automatyczne ograniczenie przechowywanych danych meta może być także sprawdzenie preferencji i ustawień dla aplikacji bądź urządzenia, którego używamy.

2. Google Hacking jako szary wywiad

Przez szary wywiad rozumie się: uzyskiwanie dostępu do treści pozostawionych nieświadomie, odtwarzanie struktury witryny lub odtwarzanie struktury sieci

[20] Maltego, https://www.paterva.com/web7/buy/maltego-clients/maltego-ce.php [dostęp: 26 I 2018].

[21] Oryon OSINT Browser, https://sourceforge.net/projects/oryon-osint-browser/ [dostęp: 26 I 2018].

[22] W przypadku plików ms Word firma Microsoft przygotowała szczegółowy poradnik na temat minimalizacji ilości metadanych, w którym użytkownik odnajdzie pełną bazę praktycznych informacji, jak ręcznie pozbyć się potencjalnego zagrożenia. Zob. *Jak zminimalizować ilość metadanych w programie Word 2003*, https://support.microsoft.com/pl-pl/help/825576/how-to-minimize-metadata-in-word-2003 [dostęp: 5 VI 2018].

[23] Podstawowym formatem zapisu metadanych są tagi EXIF, z których odczytuje się m.in. datę i czas wykonania fotografii, ustawienia aparatu oraz pozycję GPS. Inne formaty do zapisu to: IPCT czy XMP. Program ExifCleaner występuje w kilku wersjach i jest do odnalezienia na stronie dystrybutora. Daje on możliwość oczyszczenia zdjęć ze wspomnianych wcześniej informacji za pomocą jednego kliknięcia. Por. *Zanim wgrasz wakacyjne zdjęcia do sieci*…, https://niebezpiecznik.pl/post/zanim-wgrasz-wakacyjne-zdjecia-do-sieci/ [dostęp: 5 VI 2018].

wewnętrznej oraz dostęp do parametrów konfiguracyjnych serwerów www. Są to działania nieetyczne, choć legalne.

Informacje pozostawione (nieświadomie) przez twórców i właścicieli witryn internetowych

W przypadku działań określanych jako szary wywiad wyszukiwarka Google wskazuje dostęp do zasobów, które nie powinny być widoczne dla użytkowników postronnych. Są to zasoby pozostawione nieświadomie (na przykład stare treści zachowane przez administratorów strony, wewnętrzne dokumenty i materiały firmy ulokowane na serwerze i tam pozostawione), zamieszczone dla wygody i użytku wyłącznie osób, które je zamieściły (na przykład pliki z muzyką lub filmami, prywatne fotografie). Wyszukiwanie wymienionych treści można przeprowadzić z użyciem Google na wiele sposobów. Najprostsze jest odgadywanie. Jeśli w danym katalogu znajdują się na przykład pliki 5.jpg, 8.jpg i 9.jpg, to można przewidywać, że będą także na przykład pliki 1–4, 6–7 i powyżej 9. Możemy zatem potencjalnie uzyskać dostęp do materiałów, co do których nie było intencją osoby je zamieszczającej, aby zostały udostępnione. Innym sposobem jest przeszukiwanie stron internetowych pod kątem obecności na nich określonych typów treści. Wyszukiwane mogą być pliki muzyczne, zdjęcia i filmy, książki (e-booki, audiobooki). Często bywają to pliki, które użytkownik upublicznił nieświadomie (np. muzyka na serwerze ftp przeznaczona tylko do własnego użytku). Można je uzyskać na dwa sposoby: posługując się operatorem „filetype:" albo operatorem „inurl:", na przykład:

filetype:doc site:edu.pl
site:www.inp.uw.edu.pl filetype:pdf
site: www.inp.uw.edu.pl intitle:index.of.mp3

Przedmiotem pozyskania mogą być również pliki z programami. Wówczas zapytanie formułuje się w sposób następujący:

filetype:ISO

Informacje o strukturze witryn internetowych

Uzyskiwanie informacji o strukturze witryn internetowych to działanie legalne, choć wydaje się nieetyczne. Zagląda się na daną stronę internetową niejako „od kuchni", ujawnia całość jej konstrukcji, a więc postępuje się niezgodnie z intencjami jej twórców. Można to zrobić w prosty sposób, wyłącznie używając operatora „site:". Przeanalizujmy następujące wyrażenie:

site: www.inp.uw.edu.pl inp

W domenie „www.inp.uw.edu.pl" poleca się wyszukanie słowa „inp". Każda ze stron tej domeny (Google wyszukuje zarówno w tekście, jak i w tytułach oraz w nagłówku strony) zawiera to słowo, uzyskujemy więc strukturę wszystkich stron domeny. Bardziej precyzyjny wynik (choć nie zawsze możliwy do uzyskania), gdy struktura katalogów stanie się dostępna, otrzymuje się, po zastosowaniu pytania:

site:uw.edu.pl intitle:index.of "parent directory"

Odsłania ono najsłabiej zabezpieczone subdomeny uw.edu.pl, niekiedy z możliwością przeglądania całego drzewa katalogowego i pobrania wszystkich plików. Tego zapytania nie stosuje się jednak do wszystkich domen, z racji ich zabezpieczenia lub działania pod kontrolą innego serwera www.

Parametry konfiguracyjne serwerów www

Pozyskiwanie parametrów konfiguracyjnych serwerów www jest działaniem lokującym się na pograniczu szarego i czarnego wywiadu. Może ono stanowić przygotowania do gromadzenia informacji dotyczących ataku lub jedynie pobranie informacji na temat rodzaju i jakości usług świadczonych przez dane urządzenie. W celu pozyskania nazwy serwera, jego wersji oraz innych parametrów (np. portów) formułuje się zapytanie:

site:uw.edu.pl intitle:index.of server.at

Każdy z serwerów www ma swoje unikatowe frazy, np. Internet Information Service (IIS) Microsoftu:

intitle:welcome.to intitle:internet IIS

Rozpoznanie serwera www zależy wyłącznie od pomysłowości. Można na przykład spróbować tego dokonać za pomocą zapytania o jego specyfikację techniczną, podręcznik (ang. *manual*) lub tzw. strony pomocy (ang. *help pages*). Temu celowi służy zapytanie sformułowane następująco:

site:uw.edu.pl inurl:manual apache directives modules (Apache)

Dostęp do serwera może być bardziej zaawansowany – na przykład dzięki plikowi z błędami SQL:

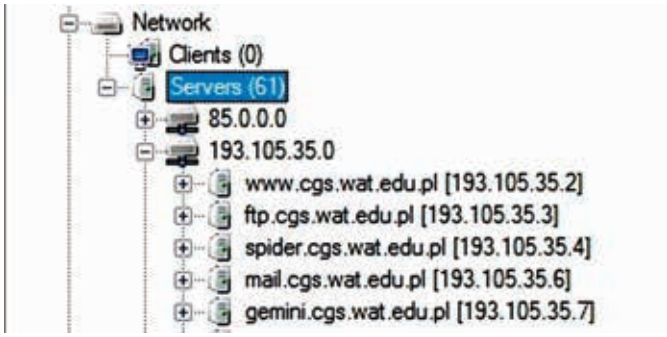
„#mysql dump" filetype:SQL

Zapisane błędy bazy SQL mogą między innymi zdradzić informacje o strukturze i zawartości bazy danych. Z kolei całość witryny, jej pierwotne i (lub) aktualne wersje można pozyskać w sposób następujący:

site:uw.edu.pl inurl:backup
site:uw.edu.pl inurl:backup intitle:index.of inurl:admin

Obecnie zastosowanie wyżej wymienionych fraz rzadko daje oczekiwane efekty, gdyż są one blokowane przez użytkowników świadomych niebezpieczeństwa. Te działania należy lokować pomiędzy szarym i czarnym wywiadem.

Za pomocą programu FOCA można również odnaleźć treści, które należy sklasyfikować w tej kategorii. Jedną z pierwszych czynności, jaką wykonuje program po rozpoczęciu pracy nad nowym projektem, jest dokonanie analizy struktury domeny oraz wszelkich subdomen zawieszonych na serwerach danej instytucji. Takie informacje można odnaleźć w okienku dialogowym, w zakładce Network.



Schemat. Zrzut ekranowy działania programu FOCA.

Źródło: Opracowanie własne.

W ten sposób potencjalny „agresor" może przechwycić treści pozostawione przez administratorów strony, a także wewnętrzne dokumenty i materiały firmy zamieszczone nawet na ukrytym serwerze.

3. Google Hacking jako czarny wywiad

Czarny wywiad to działania o charakterze nielegalnym i często uznawane jednocześnie za nieetyczne. Składają się nań: pozyskiwanie informacji celowo zabezpieczonych przed ich pozyskaniem, dostęp do osobistych danych wrażliwych (m.in. nazw użytkowników, haseł, numerów identyfikacyjnych) oraz dostęp do parametrów konfiguracyjnych urządzeń w celu przejęcia nad nimi kontroli.

Informacje celowo zabezpieczone przed pozyskaniem

Są to przede wszystkim informacje znajdujące się na witrynach udostępniających je odpłatnie (na przykład subskrypcje „The Boston Globe", „The New York Times", a z polskich – „Rzeczpospolita"). Jednym ze sposobów darmowego pozyskiwania płatnych artykułów stało się „udawanie" przez użytkownika robota internetowego Google (tzw. Googlebota, Google Spidera, pająka sieciowego Google). Stanowi on podstawowy mechanizm indeksujący strony internetowe na potrzeby wyszukiwarki Google. Ze względu na swoją funkcję programy sieciowe rozpoznawane jako Google mogą pozwolić sobie na więcej niż zwykli użytkownicy – twórcy płatnych treści chcą je indeksować, aby powiadamiać użytkowników o ich istnieniu, ale nie chcą ich udostępniać bezpłatnie. Tym samym robot Google ma (a przynajmniej jeszcze do niedawna miał) dostęp do tego typu stron. Możliwe jest imitowanie Googlebota przez zmianę identyfikatora własnej przeglądarki. Dokonuje sie tego bezpośrednio – przez zmianę wpisów konfiguracyjnych w edytorze rejestru systemu Windows, pośrednio – przez zmianę parametrów konfiguracyjnych przeglądarki albo w najprostszy możliwy sposób – przez instalację odpowiednich wtyczek do przeglądarki (dla Chrome jest to User-Agent Switcher)[24]. Pojawiły się zautomatyzowane sposoby dostępu do tych treści, używane m.in. w parze Block Referer i User Agent Switcher. Powyżej opisaną możliwość dostępu do zastrzeżonych (płatnych) treści w taki właśnie sposób zauważyli i zlikwidowali najznaczniejsi dystrybutorzy informacji (w Polsce m.in. Archiwum Rzeczpospolitej – archiwum.rp.pl – i Archiwum Gazety Wyborczej). Nieustannie jednak trwa wyścig „miecza i tarczy" – w następstwie zamknięcia tego kanału powstała między innymi wtyczka Paywall-Pass, a następnie pojawiły się mniej lub bardziej udane próby manipulacji URL artykułami archiwalnymi (np. w Archiwum Rzeczpospolitej należało zamienić końcówkę linku z „?template=restricted" na „?template=%00" i dokonać zabiegu zwanego *null byte injection* (tj. iniekcji bajtu zerowego – techniki polegającej na rozpoznaniu struktury adresu, stosowanej przez twórcę danych witryny internetowej, i takim jego przekształceniu, które spowoduje eskalację uprawnień i umożliwi wykonanie złośliwego kodu po stronie serwera). Natomiast, gdy również i tę możliwość zablokowano, nowy sposób dostępu do treści polegał na doklejeniu do URL (po html) frazy: „#ap-1" oraz „html?templateOld=00%&templateOld=0%%&". Obecnie i te możliwości zostały zablokowane, choć niewątpliwie istnieją inne. Są to jednak techniki wykraczające poza klasyczne rozumienie GH. Za pomocą samej wyszukiwarki Google informacje chronione mogą być pozyskiwane z użyciem prostych zapytań wykorzystujących błędy administratorów serwerów. Oto przykłady:

[24] Powodzenie sfałszowania swojej tożsamości można sprawdzić pod adresem: https://www.whoishostingthis.com/tools/user-agent/ [dostęp: 26 I 2018]. Warto zwrócić uwagę, że zmiana tożsamości swojej wyszukiwarki na robota sieciowego uniemożliwia korzystanie z niektórych serwisów, na przykład z poczty elektronicznej.

budżet filetype:xls
inurl:gov filetype:xls “restricted”
allintitle:sensitive filetype:doc
allintitle:restricted filetype:mail

Wrażliwe dane osobowe

Pierwszym rodzajem danych określanych jako „wrażliwe” są dane użytkowników – ich nazwy (loginy, nickname’y) wraz z hasłami. Pozyskanie tych danych umożliwia kradzież tożsamości. Tego typu informacje można nabywać za pomocą wyszukiwarki Google (i nie tylko) na wiele różnych sposobów. Są to działania ewidentnie nieetyczne i nielegalne. Powodzenie w poszukiwaniu takich informacji jest zależne od znajomości struktur systemów operacyjnych oraz struktur poszczególnych programów. Ogólnie rzecz ujmując, działania, o których mowa, polegają na tworzeniu zapytań z wykorzystaniem potencjalnych fraz i elementów współwystępujących z nazwami użytkowników i hasłami. Oto egzemplifikacja zapytania, które umożliwia odnalezienie pliku przechowującego nazwy użytkowników:

allintext:username filetype:log

Poniżej podano kilka elementarnych przykładów zapytań odnajdujących hasła użytkowników:

hasla.xlsx
passwords.xls
intitle:password
filetype:log inurl:password
intitle:”index of password”
inurl:passwd filetype:txt

W powyższych przypadkach pozyskuje się hasła pozostawione w wyniku nieodpowiedzialności administratora. Podano tu odmiany w celu ukazania licznych możliwości tego typu działań, aby unaocznić, że istotną rolę w takim poszukiwaniu, jak wyżej, odgrywa zgadywanie. Nie należy jednak liczyć na przechowywanie w ten sposób haseł przez administratorów – obecnie (2018 r.) Google zwraca zaledwie po kilka–kilkanaście wyników po użyciu takich zapytań. Poniżej – nieco bardziej zaawansowane zapytanie:

„index of /” +password.txt

Wskutek działania powyższego algorytmu są pozyskiwane hasła zapisane otwartym tekstem. Bardziej zaawansowane zapytania, z których pierwsze dotyczy haseł

administratora, a drugie wyszukuje dowolne hasła oraz inne informacje odnoszące się do logowania, wyglądają następująco:

http://admin:*@www
filetype:bak inurl:"htaccess|passwd|shadow|htusers"

Z kolei pośrednie pozyskiwanie plików haseł odbywa się w sposób przedstawiony poniżej:

inurl:config.txt

Po wpisaniu powyższej frazy uzyska się dostęp do plików config.txt, potencjalnie zawierających informacje o konfiguracji serwera, takich jak zaszyfrowane hasła administratorów lub dane umożliwiające dostęp do baz danych. Adres strony logowania można uzyskać w jeszcze prostszy sposób (w pierwszym przypadku jest to strona logowania administratora, w drugim zaś – strona zwykłych użytkowników):

inurl:adminlogin
intitle:login

Istotną rolę w pozyskiwaniu haseł odgrywa nie tylko zgadywanie, lecz także znajomość parametrów konfiguracyjnych programów. Rozważmy przykład klienta FTP (*File Transfer Protocol*). Obecnie coraz rzadziej wykorzystuje się ten rodzaj transferu przy aktualizacji, dlatego autorzy podają taką właśnie egzemplifikację. Hasła w jednym z najpopularniejszych tego typu programów (Total Commander) są zakodowane otwartym tekstem i przechowywane w pliku .ini. Zapytanie przesłane do Google w celu pozyskania haseł brzmi następująco:

inurl:wcx_ftp filetype:ini

Zasadnicze znaczenie dla pozyskiwania haseł oraz innych danych niezbędnych do autoryzacji ma analiza specyfikacji technicznej programu oraz sposobu i miejsca przechowywania tych danych w strukturze określonej aplikacji lub systemu. GH posłuży zarówno do plików zapisanych tekstem otwartym, jak i szyfrowanych. Zabezpieczone pliki można odszyfrować jednym z licznych programów służących do tego celu (popularny algorytm DES – tj. symetryczny szyfr blokowy – jest łamany przez program Jack The Ripper).

Kolejnym typem danych wrażliwych są różnego rodzaju numery identyfikacyjne, na przykład numery kart kredytowych, numery ubezpieczenia czy numery Powszechnego Elektronicznego Systemu Ewidencji Ludności (PESEL). Możliwe jest również analogiczne wyszukiwanie numerów seryjnych programów i gier. Najprostsza składnia służąca do tego celu wygląda następująco:

index of/credit-card

Podstawowym elementem umożliwiającym masowe przeszukiwanie sieci jest operator zakresu „..” oraz (fakultatywnie) operator „numrange:”:

PESEL+74010100000..76123199999
numrange:74010100000..76123199999

Powyższe zapytanie wyszukuje numery PESEL z zakresu pomiędzy 1 stycznia 1974 r. a 31 grudnia 1976 r. Wyszukanie numerów kart kredytowych odbywa się w sposób analogiczny – wystarczająca jest znajomość wzorca takiej numeracji, charakterystycznego dla konkretnego usługodawcy:

+MasterCard 5500000000000004..5599999999999999
numrange:370000000000002..379999999999999
numrange:586824160825533338..899999999999999925

Wyszukiwanie informacji o innych usługach finansowych, na przykład numerach kont, odbywa się w taki sam sposób. Aktualnie, wskutek licznych nadużyć, ta usługa jest przez Google blokowana, ale niektóre inne wyszukiwarki (na przykład Yandex) umożliwiają skorzystanie z tego sposobu wyszukiwania.

Zaawansowany GH umożliwia również pozyskiwanie innych danych wrażliwych, na przykład ze sklepów internetowych. Obecnie jednak są one już dobrze chronione. Większość pomysłów wyszukiwania pochodzi sprzed kilkunastu lat, a tylko jeden z 2016 r.[26] Wszystkie opierają się na znajomości oprogramowania takiego sklepu, na przykład:

intext:”Dumping data for table `orders`”

Powyższa fraza umożliwia odnalezienie plików zrzutu SQL, potencjalnie zawierających dane osobowe.

Parametry konfiguracyjne programów i urządzeń

Rozpoznawanie parametrów konfiguracyjnych programów i urządzeń polega na wyszukaniu luk bezpieczeństwa z zamiarem ich wykorzystania lub bezpośredniego przejęcia kontroli nad programami (urządzeniami) z użyciem fabrycznych (domyślnych) parametrów konfiguracyjnych. Również oprogramowanie serwerów rozmaitych usług

[25] Jako pierwszych użyto tzw. numerów testowych MasterCard, American Express i Maestro International.

[26] J. Long, *Google Hacking Database*, https://www.exploit-db.com/google-hacking-database/?action=search&ghdb_search_cat_id=10&ghdb_search_text= [dostęp: 26 I 2018].

w Internecie jest podatne na ataki typu GD. Cel ataku mogą stanowić systemy operacyjne, a rudymentarne zapytanie można sformułować następująco:

ip:212.85.108.185 index of /admin
ip:212.85.108.185 index of/root
ip:212.85.108.185 allinurl:winnt/system32/

Polecenia „inurl:" lub „allinurl:" umożliwiają wyszukanie maszyn zawierających określone (znane) luki bezpieczeństwa, zwłaszcza elementy konfiguracyjne, natomiast operator „ip:" wskazuje określony serwer, który jest celem ataku. W ostatnim z powyższych przykładów sprawdzamy, czy pod wskazanym adresem są katalogi systemowe dostępne przez Internet (błąd administratora). Jeśli tak, to dzięki np. plikowi cmd.exe można w następstwie przejąć kontrolę nad serwerem. Wyszukiwanie parametrów konfiguracyjnych programów może również odbywać się następująco:

inurl:config.txt
inurl:admin
filetype:cfg
inurl:server.cfg rcon password
allinurl:/provmsg.php

Istnieje możliwość wyszukania starych wersji skryptów, obsługujących takie usługi, jak fora internetowe, platformy blogowe, sklepy internetowe oraz inne serwisy, w tym klasyczne strony www. Odnalezienie nowej luki jest punktem wyjścia do poszukiwania starych wersji skryptów. Przykładem może być:

intext:"Powered by: vBulletin Version 3.7.4"

Możliwe jest również pozyskiwanie numerów seryjnych programów, w tym systemów operacyjnych. Poniżej – zapytanie GH dotyczące odnalezienia klucza rejestracyjnego Windows XP Pro:

„Windows XP Professional" 94FBR

Ataki GH nie muszą się skupiać na parametrach konfiguracyjnych programów. Przedmiot ataku mogą stanowić również ujawnione luki bezpieczeństwa. Do typowych należy zapytanie o błędy serwera (parsera). Są one często zapisywane w plikach tekstowych i niekiedy bywają indeksowane przez wyszukiwarkę Google. Typowe frazy to:

filetype:txt intext:"Access denied for user"
filetype:txt intext:"Error Message"

Google Hacking umożliwia również dostęp do rozmaitych, wyspecjalizowanych urządzeń sieciowych. Działania najpopularniejsze i najszerzej omawiane w publicystce to pozyskiwanie dostępu do kamer sieciowych. Wiele kamer jest instalowanych bez konfiguracji – bez haseł lub z loginami i hasłami fabrycznymi. Dlatego taki atak jest możliwy. Składnia zapytania, jakim należy się posłużyć, jest zależna od producenta oraz typu i rodzaju kamery. Za pomocą Google mogą być wyszukiwane ścieżki do stron logowania kamer, a także domyślne informacje powitalne. Znając te dane i umieszczając je w konstruowanym zapytaniu do wyszukiwarki, odnajdujemy liczne kamery zindeksowane z tymi danymi. Część urządzeń udostępnia intruzowi nie tylko podgląd, lecz także możliwość sterowania. Wykorzystanie operatorów typu „ip:", „site:" lub „inurl:" umożliwia zawężenie zakresu wyszukiwania do określonego podmiotu lub obszaru geograficznego. Pozyskujemy dostęp do prywatnych kamer monitorujących posesje, pokoje dziecięce lub kojce zwierząt, firmowych – monitorujących pracę biura lub jego otoczenie, a także – kamer monitoringu CCTV, urzędów i służb miejskich. Zestawienia loginów i haseł fabrycznych do kamer są powszechnie publikowane[27], istnieją również generatory haseł do kamer[28]. Z dostępem do kamer jest związany swoisty nielegalny rynek usług – są fora przeznaczone dla poszukujących i pozyskujących dostęp do tego typu wizualnych treści (między innymi anon-ib, 4chan, w mniejszym stopniu także overchan i torchan[29]). A oto przykłady zapytań, dzięki którym pozyskuje się dostęp do kamer sieciowych:

"/home/homeJ.html"
inurl:"CgiStart?page="
intitle:"Biromsoft WebCam" -4.0 -serial -ask -crack -software -a -the -build -download -v4 -3.01 -numrange:1-10000

Przedmiotem ataku mogą być również inne urządzenia, takie jak:

1) drukarki:

inurl:hp/device/this.LCDispatcher
intitle:"Dell Laser Printer" ews

2) switche:

inurl:"level/15/exec/-/show"

[27] *Hasła domyślne – Default login and password for DVR NVR IP*, kizewski.eu/it/hasla-domyslne-default-login-and-password-for-dvr-nvr-ip/ [dostęp: 26 I 2018].

[28] Patrz na przykład: http://www.cctvforum.com/viewtopic.php?f=19&t=39846&sid=42bdd50a426bea9296f1a2e78f09c226 [dostęp: 26 I 2018].

[29] *Oto jak wykrada się nagie zdjęcia gwiazd. I to nie tylko z telefonów*, https://niebezpiecznik.pl/post/oto-jak-wykrada-sie-nagie-zdjecia-gwiazd-i-to-nie-tylko-z-telefonow/, 3 IX 2014 [dostęp: 26 I 2018].

3) routery oraz inne urządzenia:

intitle:”Welcome to ZyXEL” -zyxel.com

Chcąc uzyskać nazwę użytkownika, który stworzył plik pozostawiony na serwerze danej organizacji, można skorzystać z programu FOCA. Po przeszukaniu domeny i wybraniu opcji „Download All” istnieje możliwość pobrania wszystkich odnalezionych plików, a dzięki „Extract/Analyze All Metadata” można się przekonać, czy udało się wydobyć coś interesującego. Przeszukując „Metadate Summary”, uzyskuje się informację nie tylko o nazwach użytkowników, lecz także – gdy dopisze szczęście – o ścieżkach folderów, drukarkach, zainstalowanym oprogramowaniu, adresach e-mail, systemie operacyjnym, serwerach, a być może również o hasłach, które zostały pozostawione przez nieświadomych użytkowników.

* * *

Zagrożenia generowane przez GH wynikają przede wszystkim z nieświadomości lub niefrasobliwości właścicieli i użytkowników rozmaitych programów, serwerów oraz innych urządzeń sieciowych, tak więc reguły samoobrony i ochrony informacji nie nastręczają szczególnych trudności. Zagrożenia informacyjne są związane z działaniami robota Google i robotów innych wyszukiwarek internetowych, zatem rudymentarna zasada likwidacji zagrożeń będzie się odnosić do ograniczenia lub zablokowania tym robotom dostępu do informacji zamieszczanych w sieci. Całkowitej blokady serwera www przed wyszukiwaniem przez pająki sieciowe dokonuje się przez zamieszczenie prostego pliku tekstowego zatytułowanego „robots.txt” w katalogu głównym witryny internetowej. W tym pliku należy umieścić dwie następujące komendy:

User-agent: *
Disallow: /

Umieszczenie gwiazdki eliminuje wszystkie wyszukiwarki, choć można wskazać konkretną, na przykład wpisując jej nazwę. Z kolei parametr „Disallow:” określa, które z elementów struktury strony mają zostać wyeliminowane[30]. Barierę dla pająków sieciowych można ustanawiać również na poziomie poszczególnych stron – zarówno typowych witryn internetowych, jak i blogów, a także stron konfiguracyjnych urządzeń sieciowych. W nagłówku HTML pliku należy je zaopatrzyć w jedną z następujących fraz:

<meta name=”Robots” content=”none” />
<meta name=”Robots” content=”noindex, nofollow” />

[30] Google udostępnia szczegółowe instrukcje pod adresem: http://www.google.com/remove.html [dostęp: 26 I 2018].

Jeśli powyższy zapis, w dowolnej z dwóch wersji, umieści się na stronie głównej, wówczas żadna ze stron podrzędnych, a także strona główna, nie będą indeksowane przez robota Google. Pozostaną więc bezpieczne przed GH. Tę frazę można umieścić również na stronach, które mają być omijane przez Googlebota. Jest to jednak rozwiązanie, którego bezpieczeństwo opiera się na zasadach dżentelmeńskich. Niezależnie od tego, że pająki sieciowe Google i innych wyszukiwarek honorują powyższe ograniczenia, to istnieją roboty sieciowe „polujące" na tego typu zapisy, aby pozyskać dane, które nie mają być indeksowane. Spośród bardziej zaawansowanych sposobów ochrony wypada zaproponować system CAPTCHA (*Completely Automated Public Turing test to tell Computers and Humans Apart*), tj. zabezpieczenie mające na celu dopuszczenie do skorzystania z treści strony wyłącznie człowieka, a nie tworów wirtualnych pobierających treści w sposób automatyczny. To rozwiązanie ma jednak wady – jest uciążliwe dla użytkowników takich informacji. Możliwości samoobrony zwiększają się wraz ze zdobywaniem przez administratora wiedzy i nabywaniem przez niego świadomości na temat technik GH. Prostą metodą samoobrony ograniczającą Google Dorks może być na przykład kodowanie znaków w plikach administracyjnych, utrudniające (choć nie uniemożliwiające) stosowanie GH za pomocą kodów ASCII.

Potentat na rynku bezpieczeństwa informacyjnego – McAfee – rekomenduje sześć następujących praktyk, które administratorzy powinni stosować, aby uniknąć niebezpieczeństw związanych z GH:

1) systematyczne uaktualnianie systemów operacyjnych, usług i aplikacji,
2) wdrożenie i utrzymywanie systemu zabezpieczeń zapobiegających włamaniom,
3) uwzględnianie działania robotów i wyszukiwarek, nabywanie wiedzy, co może być upubliczniane wskutek GH i jak to zweryfikować,
4) usuwanie wrażliwych zasobów z lokalizacji publicznych,
5) konsekwentny podział na to, co jest dostępne publicznie, i na to, co pozostaje prywatne. W rezultacie tego – blokowanie dostępu do zasobów nieprzeznaczonych dla użytkowników zewnętrznych,
6) częste wykonywanie testów penetracyjnych (pentestów)[31].

Szczególnie istotna wydaje się ostatnia z wymienionych praktyk, testy penetracyjne bowiem jednoznacznie określają poziom podatności witryny lub serwera na zagrożenia, w tym na GH. Istnieją narzędzia służące do przeprowadzania pentestów w obszarze GH. Jednym z nich jest Site Digger (w wersji 3.0 niewymagający odpłatnej licencji Google API)[32], umożliwiający zautomatyzowane testowanie bazy zapytań Google Hacking Data Base na dowolnie wybranej stronie internetowej. Takich narzędzi jest więcej: dość, by wspomnieć o skanerze Wikto[33] czy skanerach online[34].

[31] C. Woodward, *Go Dork Yourself! Because hackers are already dorking you*, https://securingtomorrow.mcafee.com/business/google-dorking/ [dostęp: 26 I 2018].

[32] McAfee, *SiteDigger v3.0 Released 12/01/2009*, https://www.mcafee.com/uk/downloads/freetools/sitedigger.aspx [dostęp: 26 I 2018].

[33] Sectools.org, *Wikto*, http://sectools.org/tool/wikto/ [dostęp: 26 I 2018].

[34] PentestTools.com, *Google hacking,* https://pentest-tools.com/information-gathering/google-

Działają one na podobnej zasadzie. Istnieją również narzędzia ofensywne imitujące środowisko witryny internetowej, jego błędy, luki i podatności na niebezpieczeństwa, żeby atakującego zwabić, a następnie pozyskać na jego temat informacje umożliwiające przeciwdziałanie – jest to na przykład Google Hack Honeypot[35]. Zwykły, nieprzeszkolony i nieświadomy użytkownik dysponuje ograniczonymi możliwościami i narzędziami samoobrony przed Google Hacking. Przede wszystkim może wykorzystać wobec samego siebie narzędzia GH, aby przekonać się, czy i jakie informacje wrażliwe na jego temat są dostępne w sieci publicznej. Warto również cyklicznie sprawdzać takie bazy, jak: Have I been pwned[36]? oraz We Leak Info, w celu przekonania się, czy zabezpieczenia naszych kont w Internecie nie zostały przełamane i upublicznione. Pierwsza z wymienionych baz jest dostępna pod adresem https://haveibeenpwned.com/ i obejmuje te strony, na których znalazły się dane naszych kont (na przykład: adres e-mail, login do serwisu, hasło, inne dane), wskutek tego, że te strony były zbyt słabo zabezpieczone. Wyszukiwanie prowadzi się przez wpisanie adresu poczty elektronicznej. Aktualnie (czerwiec 2018 r.) baza zawiera ponad pięć miliardów kont. Bardziej zaawansowane narzędzie znajduje się pod adresem https://weleakinfo.com. Pozwala ono na wyszukiwanie informacji po nazwie użytkownika, adresie poczty elektronicznej, haśle i jego skrócie (ang. *hash*), adresie IP, nazwisku oraz numerze telefonu. Jednak usługa wyszukiwania nie jest tu usługą jedyną – konta, których dane wyciekły, mogą być w serwisie zakupione. Jednodniowy dostęp to wydatek zaledwie dwóch dolarów.

Omówienie sposobów zapobiegania wyciekowi informacji przez GH/GD w istotnym stopniu przekracza możliwości objętościowe niniejszego artykułu. Warto jednak zaznaczyć, że profesjonalna ochrona jest możliwa jedynie po audycie eksperckim przeprowadzonym przez tzw. pentesterów lub bughunterów. Testy penetracyjne (ang. *penetration tests*, *pentests*) stają się obecnie ważną subdyscypliną praktycznego zapewniania bezpieczeństwa informatycznego oraz informacyjnego. Przeprowadza się je wielowymiarowo, testując bezpieczeństwo informatyczne, techniczne i fizyczne urządzeń oraz obiektów. Powstały specjalne narzędzia informatyczne (na przykład system operacyjny Kali Linux czy program Metasploit), wdrożono również standardy ich wprowadzania. Korzysta z nich zarówno sektor prywatny (przedsiębiorstwa), jak i publiczny. Są one jednak stosunkowo kosztowne.

O skali i randze zagrożenia mogą świadczyć informacje publikowane przez Federalne Biuro Śledcze USA (FBI), które rocznie odnotowuje kilkadziesiąt tysięcy tego typu incydentów[37]. Szacuje się ponadto, że ponad trzy czwarte stron www ma

hacking [dostęp: 31 I 2018].

[35] The Google Hack Honeypot, http://ghh.sourceforge.net/ [dostęp: 26 I 2018].

[36] Słowo „*pwned*" pochodzi z socjolektu graczy online (po raz pierwszy zostało użyte w Warcraft). Powstało wskutek błędu językowego – zamiast słowa „*owned*" (litery „p" i „o" sąsiadują na klawiaturze QWERTY). Oznacza ono dokładnie to samo, co „*to own*" – 'posiąść', jednak z naddatkiem symbolicznym, że odnosi się do sfery online.

[37] Zob.: *Roll Call Release. Intelligence for Police, Fire, EMS, and Security Personnel,* 7 lipca 2014, https://info.publicintelligence.net/DHS-FBI-NCTC-GoogleDorking.pdf [dostęp: 26 I 2018].

tzw. luki bezpieczeństwa, natomiast co dziesiąta ze stron – luki poważne[38]. Google Hacking stanowi istotne zagrożenie bezpieczeństwa informacyjnego. Koszty pozyskania danych w ten sposób są wielokrotnie mniejsze niż w przypadku ataku za pomocą innych metod[39]. Bariera kompetencji, którą trzeba pokonać, nie jest zbyt duża – jak wskazywał sam twórca terminu Google Hacking – istotą tych działań, podobnie jak innych działań hakerskich, jest prostota[40]; nie wymagają one kwalifikacji informatycznych, a jedynie zrozumienia i zapamiętania kilkunastu komend oraz pewnej dozy logicznego, a jednocześnie kreatywnego myślenia.

Bibliografia

Caballero J., *Detection of Intrusions and Malware, and Vulnerability 2016*, San Sebastian 2016, Springer.

Cucu P., *How Malicious Websites Infect You in Unexpected Ways. And what you can do to prevent that from happe*ning, „Heimdal Security", https://heimdalsecurity.com/blog/malicious-websites/ [dostęp: 26 I 2018].

Długosz M., *Legalny hacking w Polsce. (Analiza)*, http://www.cyberdefence24.pl/legalny-hacking-w-polsce-analiza [dostęp: 5 VI 2018].

Glijer C., *Ranking światowych wyszukiwarek 2017: Google, Bing, Yahoo, Baidu, Yandex, Seznam*, „K2 Search", http://k2search.pl/ranking-swiatowych-wyszukiwarek-google-bing-yahoo-baidu-yandex-seznam/ [dostęp: 26 I 2018].

Johnson L.K., *National Security Intelligence*, Oxford 2010, Oxford University Press.

Kassner M., *Google hacking*: *It's all about the dorks*, https://www.techrepublic.com/blog/it-security/google-hacking-its-all-about-the-dorks/ [dostęp: 26 I 2018].

Laskowski M., *Using Google search engine to get unauthorized access to private data,* „Actual Problems of Economics" 2012, nr 132, s. 381–386.

Long J., *Google Hacking Database*, https://www.exploit-db.com/google-hacking-database/?action=search&ghdb_search_cat_id=10&ghdb_search_text= [dostęp: 26 I 2018].

Long J., *Google Hacking for Penetration Testers*, Rockland 2007, Syngress Publishing Inc.

[38] Zob.: P. Cucu, *How Malicious Websites Infect You in Unexpected Ways. And what you can do to prevent that from happe*ning, „Heimdal Security" z 30 czerwca 2017, https://heimdalsecurity.com/blog/malicious-websites/ [dostęp: 26 I 2018].

[39] M. Laskowski, *Using Google search engine to get unauthorized access to private data*, „Actual Problems of Economics" 2012, nr 132, s. 381–386.

[40] M. Kassner, *Google hacking: It's all about the dorks*, „IT Security", https://www.techrepublic.com/blog/it-security/google-hacking-its-all-about-the-dorks/ [dostęp: 26 I 2018].

Long J., *The Google Hacker's Guide. Understanding and Defending Against the Google Hacker*, http://pdf.textfiles.com/security/googlehackers.pdf [dostęp: 26 I 2018].

Long J., Gardnem B., Brown J, *Google Hacking for Penetration Testers*, Amsterdam i in. 2005, Elsevier.

Radoniewicz F., *Odpowiedzialność karna za przestępstwo hackingu*, https://www.iws.org.pl/pliki/files/Filip%20Radoniewicz%2C%20Odpowiedzialno%C5%9B%C4%87%20karna%20za%20przest%C4%99pstwo%20hackingu%20%20121.pdf [dostęp: 5 VI 2018].

Roll Call Release. Intelligence for Police, Fire, EMS, and Security Personnel, https://info.publicintelligence.net/DHS-FBI-NCTC-GoogleDorking.pdf [dostęp: 26 I 2018].

Smart searching with googleDorking, „Exposing the Invisible", https://exposingtheinvisible.org/guides/google-dorking/#dorking-operators-across-google-duckduckgo-yahoo-and-bing [dostęp: 26 I 2018].

Turaliński K, *Wywiad gospodarczy i polityczny. Podręcznik dla specjalistów ds. bezpieczeństwa, detektywów i doradców gospodarczych*, Warszawa 2015, ARTEFAKT.edu.pl sp. z o.o.

Woodward C., *Go Dork Yourself! Because hackers are already dorking you*, https://securingtomorrow.mcafee.com/business/google-dorking/ [dostęp: 26 I 2018].

Źródła internetowe

http://ghh.sourceforge.net.

http://sectools.org/tool/wikto.

http://www.cctvforum.com/viewtopic.php?f=19&t=39846&sid=42bdd50a426bea9296f1a2e78f09c226.

https://haker.edu.pl/2015/07/10/google-hacking-google-dorks.

https://kizewski.eu/it/hasla-domyslne-default-login-and-password-for-dvr-nvr-ip.

https://niebezpiecznik.pl/post/oto-jak-wykrada-sie-nagie-zdjecia-gwiazd-i-to-nie-tylko-z-telefonow.

https://niebezpiecznik.pl/post/zanim-wgrasz-wakacyjne-zdjecia-do-sieci.

https://sourceforge.net/projects/oryon-osint-browser.

https://support.google.com/webmasters/answer/35287?hl=en.

https://support.microsoft.com/pl-pl/help/825576/how-to-minimize-metadata-in-word-2003.

https://www.collinsdictionary.com/submission/9695/google+dorks.

https://www.elevenpaths.com/labstools/foca/index.html.

https://www.mcafee.com/uk/downloads/free-tools/sitedigger.aspx.

https://www.paterva.com/web7/buy/maltego-clients/maltego-ce.php.

https://www.whoishostingthis.com/tools/user-agent.

Abstrakt

W artykule dokonano analizy potencjału technik pozyskiwania informacji w Internecie określanych mianem Google Hacking (GH), tj. formułowania zapytań dla przeglądarki Google, ujawniających dane niedostępne bezpośrednio lub te, których pozyskanie jest nieuprawnione z powodów etycznych, prawnych lub z obu tych przyczyn. Techniki zdobywania informacji metodą GH pogrupowano w trzy zbiory. Pierwszy sposób pozyskiwania danych, który nie budzi zastrzeżeń etycznych i prawnych, określono mianem białego wywiadu. Zaliczono do niego wyszukiwanie stron usuniętych i archiwalnych, wyszukiwanie niektórych informacji o użytkownikach oraz innych informacji merytorycznych. Do drugiej grupy technik, określanych jako szary wywiad, wątpliwych z etycznego punktu widzenia, włączono zdobywanie informacji pozostawionych nieświadomie przez twórców i właścicieli witryn internetowych, informacji o strukturze witryn internetowych oraz parametrów konfiguracyjnych serwerów www. Ostatnią grupę technik stanowi czarny wywiad, tj. działania nielegalne i najczęściej nieetyczne. Przeanalizowano tu potencjał uzyskiwania informacji zabezpieczonych, osobowych danych wrażliwych oraz parametrów konfiguracyjnych programów i urządzeń. Uzupełnienie tekstu stanowi analiza możliwości pozyskiwania informacji za pomocą programu FOCA (*Fingerprinting Organizations with Collected Archives*), służącego do automatyzowania zapytań GH i zorientowanego na *metadata harvesting*, czyli masowe odnajdywanie i analizowanie metadanych zawartych w dokumentach online.

Słowa kluczowe: Google Hacking, FOCA, *metadata harvesting*, wyszukiwarka internetowa.