

WOJSKOWA AKADEMIA TECHNICZNA

im. Jarosława Dąbrowskiego

WYDZIAŁ CYBERNETYKI



PRACA DYPLOMOWA

STACJONARNE STUDIA I°

Temat pracy: **METODY CYFROWEGO ODCISKU PALCA
PRZEGLĄDAREK INTERNETOWYCH I
URZĄDZEŃ PODŁĄCZONYCH DO INTER-
NETU—WYZWANIA I ROZWIĄZANIA**

INFORMATYKA W MEDYCYNIE

.....
(kierunek studiów)

INFORMATYCZNE SYSTEMY ZARZĄDZANIA W MEDYCYNIE

.....
(specjalność)

Dyplomant:

Artur WOLFF

Promotor pracy:

dr inż. Rafał KASPRZYK

Warszawa 2021

Oświadczenie

Wyrażam zgodę na udostępnianie mojej pracy w czytelni Archiwum WAT.

Dnia

Pracę przyjąłem

promotor pracy
dr inż. Rafał Kasprzyk

Spis treści

1. Wprowadzenie do fingerprintingu	7
1.1. Podstawowe pojęcia	7
1.1.1. Nomenklatura używana w tej pracy	7
1.1.2. Definicje	8
1.1.3. Właściwości fingerprintu	9
1.2. Fingerprinting a Internet	10
1.2.1. Początki Internetu	10
1.2.2. Założenia funkcjonowania Internetu i ich realizacja	11
1.2.3. Początki śledzenia użytkowników Internetu	11
1.3. Fingerprinting w branży komputerowej	15
1.3.1. Fingerprinting audio, wideo i algorytmy ACR	15
1.3.2. Fingerprinting klucza publicznego	15
2. Problematyka prywatności i anonimowości w Internecie	17
2.1. Metody śledzenia użytkowników Internetu	17
2.2. Zagrożenia związane z fingerprintingiem	17
2.3. Możliwości ochrony prywatności	17
2.4. Pozytywne aspekty fingerprintingu	17
3. Metody fingerprintingu urządzeń i przeglądarek	18
3.1. Pojęcie pasywnego, pół pasywnego i aktywnego fingerprintingu	18
3.2. Źródła danych identyfikacyjnych urządzeń	18
3.2.1. Protokoły warstwy drugiej w modelu OSI	18
3.2.2. Stos TCP/IP	18

3.2.3. Nierutowalne protokoły warstwy 5 (lokalny fingerprinting) . . .	19
3.2.4. Rutowalne protokoły warstwy 7	19
3.3. Wybrane metody fingerprintingu urządzeń i ich systemów operacyjnych	19
3.4. Przeglądarki jako specjalny przypadek fingerprintingu urządzeń	19
3.5. Źródła danych identyfikacyjnych przeglądarek	19
3.6. Wybrane metody fingerprintingu przeglądarek	19
3.6.1. Implementacje wybranych metod fingerprintingu przeglądarek	19
3.7. Implementacja przykładu identyfikacji przeglądarki	19
4. Eksperymentalny algorytm klasyfikatora fingerprintów	20
4.1. Motywacja za stosowaniem klasyfikatora fingerprintów	20
4.2. Projekt i implementacja algorytmu	20
4.3. Ocena złożoności czasowej i pamięciowej algorytmu	20
4.4. Ocena efektywności algorytmu w kontekście klasyfikatora	20
5. Podsumowanie	21

Wstep

Wstep

Rozdział 1.

Wprowadzenie do fingerprintingu

1.1. Podstawowe pojęcia

1.1.1. Nomenklatura używana w tej pracy

Pisząc o odcisku palca użyto (także w tytule pracy) ogólnie przyjętego skrótu myślowego, oznaczającego odbitkę linii papilarnych, czyli formę językową uznawaną za poprawną przez specjalistów od daktyloskopii.

Użycie formy językowej „odcisk palca” w terminie „cyfrowy odcisk palca” ma wiele sensu. Jeszcze bez zdefiniowania tego specjalistycznego terminu, możemy domyślić się, co oznacza. Oczywiście wynika to z faktu, że cyfrowy odcisk palca i analogowy odcisk palca są ze sobą w pewien sposób powiązane (koncepcja cyfrowego odcisku palca czerpie z wartości wynikających ze stosowania odbitek ludzkich linii papilarnych w dziedzinie kryminalistyki).

Angielskie słowo „fingerprint” tłumaczy się jako odcisk palca, jednakże w zagranicznych publikacjach dotyczących cyfrowego odcisku palca rzadko występuje termin „digital fingerprint”. Kontekst użycia jest na tyle wyraźny, że użycie samego „fingerprint” jest wystarczające.

Zachodnie nazewnictwo ma tę przewagę, że jest zdecydowanie bardziej kompaktowe. Także w przypadku słotwórczego zabiegu *fingerprinting*, oznaczającego czynność; szukając polskiego odpowiednika musielibyśmy sięgnąć po „cyfrowe znakowanie”. Z uwagi na tę kompaktowość i łatwość użycia, w pracy pre-

ferowane będzie użycie oryginalnej nomenklatury.

1.1.2. Definicje

W kolejnych punktach zawarto najważniejsze definicje i powiązane pojęcia, które będą używane w przeciągu całej pracy.

Fingerprint

Wektor cech pozwalający zidentyfikować dowolny zbiór danych.

Aby *fingerprint* pełnił praktyczną funkcję identyfikacyjną, tak jak odfisk ludzkich linii papilarnych pełni praktyczną funkcję identyfikacyjną, często stosuje się algorytm, który kojarzy wektor cech z określonej długości (zwykle krótkim) ciągiem bajtów (identyfikatorem; można go także rozumieć jako etykieta). Takim algorytmem może być na przykład wysokiej wydajności funkcja skrótu (niekoniecznie zdalna do zastosowań kryptograficznych—na przykład MurmurHash). W niektórych źródłach można także spotkać się z taką definicją, że *fingerprint* to już sam wynik wyżej wspomnianego algorytmu. Taka definicja nie zmienia istoty *fingerprintu*, ale jest zdecydowanie mniej przydatna w kontekście *fingerprintingu* urządzeń podłączonych do Internetu i przeglądarek internetowych, czego dotyczy niniejsza praca.

Fingerprint urządzenia podłączonego do Internetu

Wektor cech pozwalający zidentyfikować urządzenie podłączone do Internetu.

Instalacja przeglądarki internetowej

Instalacja na konkretnym urządzeniu. W przypadku zmiany ustawień, konfiguracji i liczby pluginów oraz aktualizacji przeglądarki, instalacja przeglądarki pozostaje ciągle tą samą instalacją.

Fingerprint przeglądarki internetowej

Wektor cech pozwalający zidentyfikować instalację przeglądarki internetowej.

1.1.3. Właściwości fingerprintu

Ludzkie linie papilarne są na ogół niepowtarzalne, niezmiennie i nieusuwalne. Z wartości wynikających ze stosowania ich odbitek w swojej dziedzinie badawczej czerpie (także etymologicznie) koncepcja *fingerprintu* i dlatego też *fingerprint* z dobrze dobranymi cechami będzie odzwierciedlać podobne właściwości.

W przypadku *fingerprintingu* urządzeń podłączonych do Internetu i przeglądarek internetowych najważniejszymi ich właściwościami są unikalność / różnorodność (niepowtarzalność) oraz stabilność (niezmiennność), przy czym zwiększenie unikalności lub stabilności ma najczęściej negatywny wpływ na drugi parametr.

Jedną ze stosowanych¹ metod pomiaru unikalności *fingerprintu* urządzeń i przeglądarek jest entropia Shannona.

Entropia Shannona

Wartość entropii można rozumieć jako liczbę pytań binarnych potrzebnych do sklasyfikowania losowo wybranego elementu z danego zbioru. Zatem entropia Shannona zbioru D z etykietami $\{l_0, l_1, \dots, l_{n-1}\}$ wyraża się wzorem

$$H(D) = - \sum_{i=0}^{n-1} p(l_i) \log_2 p(l_i)$$

gdzie $p(l_i)$ to wyrażona ułamkiem częstość $x \in D$ mającego etykietę l_i . W przypadku w którym każda etykieta występuje tak samo często entropia ma wartość maksymalną równą $\log_2 n$.

Przykład Jeśli zbiór *fingerprintów* przeglądarek internetowych ma 32 bity entropii, to w przypadku losowego wyboru jednego z nich oczekujemy, że w najlepszym przypadku tylko 1 na 4294967295 przeglądarek będzie miała taki sam *fingerprint*.

¹Metrykę tę stosowało na przykład badanie „How Unique Is Your Web Browser?” Electronic Frontier Foundation; w momencie pisania pracy jedno z największych badań tego typu.

Stabilność

W przypadku dodania kolejnej cechy do wektora cech, identyfikującego urządzenie lub przeglądarkę, zwykle zwiększy to entropię, ale także zmniejszy stabilność *fingerprintu*. Dzieje się tak ponieważ jest to kolejna rzecz, która może zmienić się w czasie. Jeśli jedną z cech wejściowych jest wersja oprogramowania urządzenia lub wersja przeglądarki (która zwykle zmienia się parę razy w ciągu roku), to kolejne *fingerprints* mogą odbiegać od siebie i naiwny klasyfikator, korzystający z algorytmu reagującego na najmniejsze zmiany (na przykład funkcja skrótu), mógłby nadać takiemu urządzeniu/przeglądarce kolejną etykietę, zamiast potraktowania jej jako poprzednio widzianą instalację.

1.2. Fingerprinting a Internet

Aby lepiej zrozumieć istotę *fingerprintu* i motywację stojącą za stosowaniem *fingerprintingu* w kontekście urządzeń podłączonych do Internetu oraz przeglądarek internetowych, wspominając o różnych innych obszarach przetwarzania komputerowego w których wykorzystywany jest *fingerprinting* w stosownych mu celach, kolejne punkty posłużą jako referencja (także historyczna).

1.2.1. Początki Internetu

Początek Internetu jaki znamy obecnie, to początek stworzonej w 1969 roku na potrzeby amerykańskiego wojska sieci ARPANET. ARPANET była implementacją koncepcji rozproszonych sieci cyfrowych transmisji danych Paula Barana z 1962 roku. Na samym początku swojego istnienia, Internet wykorzystywany był do tego aby rozpraszać obliczenia pomiędzy wiele komputerów—w tym wypadku chodziło o superkomputery znajdujące się w innych ośrodkach badawczych (ARPANET powstało na Uniwersytecie Kalifornijskim w Los Angeles). W tym samym okresie powstawały inne globalne sieci komputerowe, zapoczątkowane zwykle w innym celu (na przykład komunikacyjnym, rozrywkowym), które później połączono z ARPANET. Ba-

dacze historii Internetu wskazują na fakt, iż gwałtowny rozwój Internetu zawdzięcza się właśnie komunikacyjnemu i rozrywkowemu aspektowi konkurencyjnych sieci.

1.2.2. Założenia funkcjonowania Internetu i ich realizacja

Po tym jak w 1989 Tim Berners-Lee oraz Robert Cailliau utworzyli projekt sieci dokumentów hipertekstowych, czyli tego, co obecnie znamy jako World Wide Web i strony internetowe, osoby prywatne oraz instytucje komercyjne zaczęły dostrzegać korzyści z użytkowania Internetu, a szczególnie z wykorzystania go jako medium reklamy i sprzedaży. Zniesienie zakazu wykorzystywania Internetu do celów zarobkowych w 1991 roku zakończyło chwilę w której Internet był medium naukowego dyskursu i zapoczątkowało okres istnienia Internetu dla mas, który trwa do dziś.

Perspektywa techniczna

Podstawą struktury obecnego Internetu jest model TCP/IP i koncepcyjnie składa się ze współpracujących ze sobą 4 warstw:

1. dostępu do sieci
2. kontroli transportu
3. Internetu
4. aplikacji

W najwyższej z warstw, czyli warstwie aplikacji działają takie usługi jak przeglądarka czy serwer WWW. To najbardziej interesująca warstwa z perspektywy niniejszej pracy, ale *fingerprinting* urządzeń podłączonych do Internetu może odbywać się także w niższych warstwach.

1.2.3. Początki śledzenia użytkowników Internetu

Gwałtowny rozwój komercyjnego Internetu sprawił, że firmy zajmujące się reklamą i sprzedażą w Internecie zaczęły także dostrzegać korzyści płynące ze śledzenia

użytkowników Internetu. W szczególności zaczęto analizować aktywność i zachowanie użytkowników. Oprócz instytucji komercyjnych, śledzeniem użytkowników zainteresowane są instytucje rządowe, co dobitnie pokazał wyciek poufnych, tajnych i ściśle tajnych dokumentów NSA w 2013 roku. Techniki śledzenia zmieniały się w czasie, wraz z rozwojem Internetu.

Adres IPv4

Adres IPv4 na początku istnienia Internetu był swego rodzaju globalnym identyfikatorem, dzięki któremu można było unikatowo identyfikować użytkowników Internetu. Adres IPv4 to 32 bitowy identyfikator. Prosta estymacja pozwala nam zauważyć, że adresów IP w wersji czwartej jest około 4,3 miliarda². Internet dzisiaj, to wielomiliardowa społeczność, a liczba urządzeń podłączonych do Internetu zdecydowanie przewyższa wyżej wymienioną estymację. Już w 1992 roku zauważono, że w najbliższym czasie pula adresów IPv4 zostanie wyczerpana³. W następnych latach proponowano kolejne rozwiązania (takie jak na przykład NAT⁴), które implementowali dostawcy usług internetowych, pozwalając na łączenie się wielu urządzeń za pośrednictwem jednego, publicznego adresu IPv4. Wyczerpywanie się kolejnych pul adresów pokazuje Rys. 1.

Biorąc pod uwagę powyższe, na mocy Zasady Szufladkowej Dirichleta możemy stwierdzić, że adres IPv4 nie jest już identyfikatorem, który mógłby unikatowo identyfikować każde urządzenie podłączone do Internetu. W tym momencie warto także zaznaczyć, że o ile nowy standard IPv6 pozwalałby na taką identyfikację, to został on zaprojektowany z myślą o prywatności i posiada szereg rozszerzeń, które w przyszłości (kiedy Internet w pełni przejdzie na adresację w wersji szóstej) będą zapobiegać precyzyjnej identyfikacji.

²W rzeczywistości liczba adresów IP w wersji czwartej jest niższa.

³RFC 1338

⁴RFC 1631



Źródło: <https://upload.wikimedia.org/wikipedia/commons/c/cf/lpv4-exhaust.svg>

Rys. 1. Wolne pule adresów IPv4 w czasie

Cookies

Małe porcje informacji zapisywane na urządzeniu użytkownika, w obszarze pamięci trwałej przeglądarki, po interakcji ze stroną internetową, która je zapisuje. Powstały głównie ze względu na potrzebę poprawienia doświadczeń użytkowników ze stronami internetowymi, tak aby zapamiętywać pewien stan, o znaczeniu dla danej sesji dla danego użytkownika (na przykład stan koszyka w sklepie internetowym).

Cookies dzielą się na tak zwane *first-party cookies* i *third-party cookies*. O ile pierwsze z wymienionego podziału faktycznie desygnowane są do tego aby spełniać wymienioną funkcję, to *third-party cookies* mogą być nadawane przez skrypty reklamowe (na przykład), umieszczone na serwującej je stronie, dzięki czemu użytkownik może być śledzony w kontekście całej sieci reklamowej. Ubogie mechanizmy kontroli cookies w przeglądarkach internetowych i obawy związane z naruszaniem prywatności użytkowników przez śledzenie wykorzystujące cookies, doprowadziły do powstania dyrektywy Unii Europejskiej dotyczącej obowiązku informacyjnego, która zawiera m.in. obowiązek informowania o polityce stosowania cookies. Doprowadziło to wcześniej także do powstania rozszerzeń w przeglądarkach, takich jak nagłówki Do Not Track i Tryb Prywatny, które w domyśle miały pomóc częściowo rozwiązać wspomniany problem.

Niektóre przeglądarki internetowe, takie jak Apple Safari (Intelligent Tracking Prevention silnika WebKit) lub Mozilla Firefox wykorzystują obecnie natywne mechanizmy inteligentnego blokowania *third-party cookies*. Powstało także wiele rozszerzeń do przeglądarek, które pozwalają blokować niechciane cookies.

Inne techniki śledzenia użytkowników

Fingerprinting urządzeń podłączonych do Internetu i przeglądarek internetowych, to jedna ze zbioru wymyślnych technik, które zaczęto stosować ze względu na ułomność i/lub postępujące ograniczenia technik wykorzystujących na przykład adresy IP urządzeń i/lub cookies.

1.3. Fingerprinting w branży komputerowej

Fingerprinting to technika wykorzystywana w wielu obszarach w dyscyplinie informatyki; *fingerprinting* urządzeń podłączonych do Internetu i przeglądarek internetowych to tylko pewien wycinek zastosowań tej koncepcji. O ile podane na początku tego rozdziału definicje ujmują *fingerprint* w sposób ogólny i także taki, który jest spójny z definicjami, które można znaleźć w publikacjach dotyczących *fingerprintingu* urządzeń i przeglądarek, to definicje dotyczące zastosowań *fingerprintu* innych bytów mogą być bardziej specyficzne czy też mogą eksponować inne właściwości, charakterystyczne dla stosownych zastosowań. Kolejne punkty służą jako referencja do ukazania jak szeroko wykorzystywana jest omawiana koncepcja.

1.3.1. Fingerprinting audio, wideo i algorytmy ACR

Metody *fingerprintingu* akustycznego i cyfrowych materiałów wideo, znane także jako algorytmy Automatic Content Recognition (ACR), powstały głównie ze względu na potrzebę kontroli treści umieszczanych w serwisach internetowych pod kątem potencjalnych naruszeń praw do wykorzystania. Algorytmy Automatic Content Recognition są sparametryzowane w podobny sposób, co algorytmy *fingerprintingu* urządzeń i przeglądarek, czyli istotny jest balans pomiędzy niepowtarzalnością a stabilnością *fingerprintu*. Klasyfikacja *fingerprintu* audio lub wideo musi działać w podobny sposób do zachowania ludzkiego moderatora, czyli w przypadku kiedy materiał jest nieodróżnialny dla ludzkiego ucha lub oka jako ten, wobec którego przeprowadzany jest proces rozpoznawania, to powinien on zostać oflagowany. Proces wykorzystywany przez algorytmy ACR określany jest mianem *perceptual hashing*⁵.

1.3.2. Fingerprinting klucza publicznego

W kryptografii klucza publicznego, w celach autoryzacji klucza publicznego pozyskanego w niezaufany sposób (na przykład ściągając go ze strony internetowej),

⁵Termin nie został przetłumaczony umyślnie, ze względu na fakt, iż polskie nazewnictwo jest w tym przypadku raczej nieistniejące.

poprzez zaufany kanał wymiany informacji (np. rozmowa telefoniczna), który nie pozwala na autoryzację całego klucza w efektywny sposób, stosuje się jego skrót, nazywany „fingerprintem klucza publicznego”. Wynik zastosowania odpowiedniej funkcji skrótu na kluczu publicznym jest na tyle kompaktowy, że pozwala na efektywną manualną autoryzację, czyli ręcznie przez człowieka. W celu ułatwienia wymiany *fingerprintów* kluczy publicznych poprzez kanały głosowe powstała lista słów PGP, która analogicznie do alfabetu fonetycznego NATO asocjuje każdą kolejną porcję bitów *fingerprintu* klucza z odpowiednim słowem w języku angielskim.

Rozdział 2.

Problematyka prywatności i anonimowości w Internecie

Technika śledzenia urządzeń, przeglądarek i użytkowników o której traktuje ta praca jest „chodliwym” tematem głównie dlatego, że jej implikacje w obszarze prywatności i anonimowości są istotne.

2.1. Metody śledzenia użytkowników Internetu

2.2. Zagrożenia związane z fingerprintingiem

2.3. Możliwości ochrony prywatności

2.4. Pozytywne aspekty fingerprintingu

Rozdział 3.

Metody fingerprintingu urządzeń i przeglądarek

3.1. Pojęcie pasywnego, pół pasywnego i aktywnego fingerprintingu

3.2. Źródła danych identyfikacyjnych urządzeń

3.2.1. Protokoły warstwy drugiej w modelu OSI

3.2.2. Stos TCP/IP

Protokoły warstwy 3 i 4 w modelu OSI

IPv4

IPv6

ICMP

IEEE802.11

- 3.2.3. Nierutowalne protokoły warstwy 5 (lokalny fingerprinting)**
- 3.2.4. Rutowalne protokoły warstwy 7**
- 3.3. Wybrane metody fingerprintingu urządzeń i ich systemów operacyjnych**
- 3.4. Przeglądarki jako specjalny przypadek fingerprintingu urządzeń**
- 3.5. Źródła danych identyfikacyjnych przeglądarek**
- 3.6. Wybrane metody fingerprintingu przeglądarek**
 - 3.6.1. Implementacje wybranych metod fingerprintingu przeglądarek**
- 3.7. Implementacja przykładu identyfikacji przeglądarki**

Rozdział 4.

Eksperymentalny algorytm klasyfikatora fingerprintów

- 4.1. Motywacja za stosowaniem klasyfikatora fingerprintów**
- 4.2. Projekt i implementacja algorytmu**
- 4.3. Ocena złożoności czasowej i pamięciowej algorytmu**
- 4.4. Ocena efektywności algorytmu w kontekście klasyfikatora**

Rozdział 5.

Podsumowanie

Spis rysunków

1.	Wolne pule adresów IPv4 w czasie	13
----	--	----

Spis tablic