

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	12
1.3. Fingerprinting w branży komputerowej	15
1.3.1. Fingerprinting audio, wideo i technologia ACR	15
1.3.2. Fingerprinting klucza publicznego	16
2. Problematyka prywatności i anonimowości w Internecie	17
2.1. Prywatność i anonimowość	17
2.1.1. Łączenie danych	18
2.1.2. Czy prywatność jest nam potrzebna?	19
2.2. Metody identyfikacji użytkowników	19
2.3. Zagrożenia związane z fingerprintingiem	21
2.4. Możliwości ochrony przed fingerprintingiem	21
2.5. Pozytywne aspekty fingerprintingu	21
3. Metody fingerprintingu urządzeń i przeglądarek	22
3.1. Pojęcie pasywnego, pół pasywnego i aktywnego fingerprintingu	22

3.2. Źródła danych identyfikacyjnych urządzeń	22
3.2.1. Protokoły warstwy drugiej w modelu OSI	22
3.2.2. Stos TCP/IP	22
3.2.3. Nierutowalne protokoły warstwy 5 (lokalny fingerprinting) . . .	23
3.2.4. Rutowalne protokoły warstwy 7	23
3.3. Wybrane metody fingerprintingu urządzeń i ich systemów operacyjnych	23
3.4. Przeglądarki jako specjalny przypadek fingerprintingu urządzeń	23
3.5. Źródła danych identyfikacyjnych przeglądarek	23
3.6. Wybrane metody fingerprintingu przeglądarek	23
3.6.1. Implementacje wybranych metod fingerprintingu przeglądarek	23
3.7. Implementacja przykładu identyfikacji przeglądarki	23
4. Eksperymentalny algorytm klasyfikatora fingerprintów	24
4.1. Motywacja za stosowaniem klasyfikatora fingerprintów	24
4.2. Projekt i implementacja algorytmu	24
4.3. Ocena złożoności czasowej i pamięciowej algorytmu	24
4.4. Ocena efektywności algorytmu w kontekście klasyfikatora	24
5. Podsumowanie	25

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”. Jak piszą Flood i Karlsson [5, s. 4], kontekst użycia jest na tyle wyraźny, że użycie samego „fingerprint” jest wyczerpujące.

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 preferowane

będzie użycie oryginalnej nomenklatury.

1.1.2. Definicje

W kolejnych punktach zawarto najważniejsze definicje i powiązane pojęcia analogiczne do tych obecnych w literaturze [2, 5], 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 zdatna 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 [16, s. 123–132]. Taka definicja nie zmienia istoty *fingerprintu*, ale jest zdecydowanie mniej przydatna w kontekście *fingerprintingu* urządzeń podłączonych do Internetu i przeglądarek internetowych.

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*/rozszerzeń 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 *fingerprintu* urządzeń podłączonych do Internetu i przeglądarek internetowych najważniejszymi jego właściwościami są unikalność / różnorodność (niepowtarzalność) oraz stabilność (niezmiennność), przy czym zwiększenie unikalności lub stabilności ma najczęściej negatywny wpływ na drugi parametr [2, s. 11].

Jedną ze stosowanych metod pomiaru unikalności *fingerprintu* urządzeń i przeglądarek jest entropia Shannona [2, s. 6].

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 posiada 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*.

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 niezależnych prac Paula Barana, Donalda Daviesa i Leonarda Kleinrocka z lat 60. XX wieku. 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) [13]. W tym samym okresie powstawały inne globalne sieci komputerowe zapoczątkowane zwykle w innym celu (na przykład komunikacyjnym, rozrywkowym), które później połączono z ARPAnet.

Badacze historii Internetu wskazują na fakt, iż gwałtowny rozwój Internetu zawdzięcza się właśnie komunikacyjnemu i rozrywkowemu aspektowi konkurencyjnych sieci [9].

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 [13]. 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 [7]:

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 z identyfikacji i śledzenia użytkowników Internetu. W szczególności zaczęto analizować aktywność i zachowanie użytkowników. Oprócz instytucji komercyjnych identyfikacją i śledzeniem użytkowników zainteresowane są instytucje rządowe, co dobitnie pokazał wyciek poufnych, tajnych i ściśle tajnych dokumentów NSA w 2013 roku. Metody identyfikacji, a zarazem śledzenia użytkowników 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 miliarda¹. 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 wyczerpana [6]. W następnych latach proponowano kolejne rozwiązania (takie jak na przykład NAT [3]), 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) mają zapobiegać precyzyjnej identyfikacji [11].

¹W rzeczywistości liczba dostępnych adresów jest niższa: <https://stackoverflow.com/a/2437185>



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 (na przykład) skrypty reklamowe 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 nagłówek 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 metody identyfikacji użytkowników

Fingerprinting urządzeń podłączonych do Internetu i przeglądarek internetowych to jedna ze zbioru wymyślnych technik, które zaczęto stosować ze względu na ułomność lub postępujące ograniczenia metod wykorzystujących na przykład adresy IP urządzeń, lub *cookies*. *Fingerprinting* przeglądarek po raz pierwszy opisał Eckerley jako technikę, która pozwala serwerom WWW jednoznacznie zidentyfikować urządzenie użytkownika za pomocą informacji wysyłanych przez przeglądarkę in-

ternetową, wtedy kiedy te informacje są unikalne dla większości z nich, tworząc ich *fingerprint* [2]. Relację urządzeń i przeglądarek pomiędzy ich użytkownikami opisuje 2.2.

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. Definicje związane z zastosowaniami *fingerprintu* innych bytów mogą być bardziej specyficzne lub mogą eksponować inne, charakterystyczne właściwości. Kolejne punkty posłużą jako referencja, do ukazania jak szeroko wykorzystywana jest omawiana koncepcja.

1.3.1. Fingerprinting audio, wideo i technologia ACR

Metody *fingerprintingu* akustycznego i cyfrowych materiałów wideo znane także jako technologia Automatic Content Recognition (ACR) zostały zaprezentowane w 2012 podczas Consumer Electronics Show, pokazując, że urządzenia dostępne dla zwykłego konsumenta mogą być na tyle sprytne, by wyszukiwać informacje na podstawie kontekstu [12]. Technologia ACR sparametryzowana jest 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 technologię Automatic Content Recognition określany jest mianem *perceptual hashing*.

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) poprzez zaufany kanał wymiany informacji (zwykle 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, o której traktuje ta praca, jest ważnym tematem głównie dlatego, że wykorzystanie jej do identyfikacji użytkowników ma istotne implikacje w obszarze prywatności i anonimowości.

2.1. Prywatność i anonimowość

Prywatność internetowa to pewien podzbiór relacji pomiędzy gromadzeniem i rozpowszechnianiem danych, technologią, społecznym oczekiwaniem wobec prywatności i prawnymi oraz politycznymi problemami orbitującymi wokół tych zagadnień [10]. W celach orientacyjnych stosowanym uproszczeniem jest pogląd, że prywatność internetowa to możliwość zatrzymania związanych z własną aktywnością danych dla siebie. Anonimowością jest zatem sytuacja, w której przekazywanie takich danych jest zablokowane.

Prywatność i anonimowość użytkowników Internetu stała się zagadnieniem jeszcze przed nadejściem ery Internetu [1], pozwalając na koniec lat dziewięćdziesiątych na zaognienie dyskusji, która trwa do dzisiaj. Użytkownicy Internetu mają różne oczekiwania wobec poziomu ich prywatności w sieci. Mniej wyczuleni na punkcie prywatności użytkownicy są w stanie pójść na pewny kompromis pomiędzy

wykorzystaniem ich danych a oferowanym na tej podstawie potencjalnym usprawnieniem ich doświadczeń w Internecie (na przykład na konkretnych stronach internetowych, w kontekście sieci reklamowych lub w innych szerszych, kontekstach). Akceptują oni ryzyko zbyt szczegółowego profilowania, potencjalnych naruszeń prywatności i inwigilacji. Inni użytkownicy dążą (mniej lub bardziej) do utrzymania anonimowości takiej, jaka panowała w Internecie na początku jego istnienia [15, s. 54–69].

2.1.1. Łączenie danych

Istnieją firmy specjalizujące się w pozyskiwaniu, kupowaniu i przetwarzaniu danych użytkowników w różnych celach (zwykle reklamowych). Nie wszystkie firmy rynku danych przetwarzają dane w sposób naruszający prywatność użytkowników, ale techniki takie jak łączenie danych z różnych źródeł (bez uprzedniej anonimizacji) mogą lub będą prowadzić do nadużyć.

Bardzo sławny przypadek firmy Cambridge Analytica, która tworzyła profile psychologiczne użytkowników i używała ich do manipulacji opinią publiczną za pomocą mediów społecznościowych to jeden z przykładów firmy, która łącząc dane, poważnie naruszyła prywatność użytkowników (m.in. Facebooka).

Istnieje wiele różnych dróg, dzięki którym w Internecie i w świecie rzeczywistym użytkownicy będą nieświadomie profilowani lub inwigilowani, a łączenie danych z różnych źródeł istotnie poprawi dokładność obrazu użytkownika. Jedną z takich dróg jest używanie stron internetowych będących częścią większej sieci reklamowej, która łączy dane na przykład z portali *social media*, wyników wyszukiwań, wysyłanych formularzy, a nawet z takich źródeł jak systemów automatycznego wykrywania twarzy w sklepach stacjonarnych. Niektóre informacje, które można wnioskować po automatycznym przetworzeniu to orientacja seksualna, poglądy polityczne i religijne, rasa, historia użycia substancji psychoaktywnych, estymowany iloraz inteligencji czy osobowość [8].

2.1.2. Czy prywatność jest nam potrzebna?

Ochrona prywatności ma swoich przeciwników i zwolenników. „nie obchodzi mnie prywatność, bo nie mam nic do ukrycia” to argument przytaczany przez przeciwników ochrony prywatności. Jedną z ważnych osób, które opowiedziały się niegdyś za taką argumentacją, jest Eric Schmidt, były CEO firmy Google. Istnieje silna polaryzacja pomiędzy przeciwnikami i zwolennikami ochrony prywatności. W swojej książce autobiograficznej Edward Snowden (słynny amerykański *whistleblower*) stwierdził, że „oświadczyć, że nie obchodzi cię prywatność, bo nie masz nic do ukrycia, to mniej więcej to samo, co oświadczyć, że nie obchodzi cię wolność słowa, ponieważ nie masz nic do powiedzenia” [15]. Bruce Schneier, amerykański kryptograf i ekspert w dziedzinie bezpieczeństwa komputerowego podsumowuje, że zbyt wiele osób myśli o tym argumencie jak o wyborze pomiędzy bezpieczeństwem a prywatnością. „Prawdziwym wyborem jest wybór pomiędzy wolnością a kontrolą” [14]. Prawo do prywatności zawiera się w Powszechnej Deklaracji Praw Człowieka².

2.2. Metody identyfikacji użytkowników

Tak ja zauważono wcześniej w niniejszej pracy—wraz z komercjalizacją Internetu powstał i ewoluował szereg różnych metod identyfikacji urządzeń, przeglądarek i tym samym użytkowników. Istnieją różne zastosowania identyfikacji, ale jednym z najbardziej powszechnie omawianych (i kontrowersyjnych) jest śledzenie użytkowników (formalnie: łączenie ze sobą wielu wizyt jednego użytkownika na tej samej platformie) [4, s. 3].

Warto także zauważyć, że *fingerprinting* przeglądarek internetowych to jedna z dróg identyfikacji urządzeń podłączonych do Internetu. Jest ona jednak na tyle reprezentatywna, że często mówi się o *fingerprintingu* urządzeń jako *fingerprintingu* przeglądarek. Dzieje się tak, ponieważ dzisiejsze przeglądarki internetowe podczas interakcji z serwerami WWW mogą aktywnie lub pasywnie przekazywać zestaw danych na tyle szeroki [2], że zawiera on w sobie *fingerprint* urządzenia, na którym

²<https://www.ohchr.org/EN/UDHR/Pages/Language.aspx?LangID=pqi>

działa przeglądarka. Przeważający ogrom aktywności użytkowników wokół „przegłądania” Internetu i wiele danych „oferowanych” przez przeglądarki internetowe naturalnie sprawia, że wysiłki identyfikujące użytkowników rozważane są głównie w kontekście tychże. Niniejsza praca stara się zapewnić pewien (choć często niewi doczny) podział. W kolejnym rozdziale niniejszej pracy omówiony jest także *fingerprinting* urządzeń podłączonych do Internetu, kiedy niemożliwe jest wykorzystanie do celów identyfikacyjnych przeglądarki internetowej użytkownika.

Identyfikacja użytkowników nie zawsze jest także synonimem z identyfikacją urządzeń czy przeglądarek, ale odsetek przypadków, kiedy w dzisiejszym świecie więcej niż jedna osoba korzysta z np. jednej przeglądarki internetowej, wydaje się stosunkowo niski. Powodem dla takiej estymacji jest fakt, iż wspomniane przeglądarki, które posiadają istotny ułamek udziałów w rynku przeglądarek internetowych, posiadają mechanizmy pozwalające na ich jednoznaczną personalizację (parowanie z personalnym kontem Google w przeglądarce Google Chrome, integracja przeglądarki Safari z ekosystemem firmy Apple, parowanie z personalnym kontem Firefox w przeglądarce Firefox itd.). Co więcej, jeden użytkownik może korzystać z wielu przeglądarek, co utrudnia jednoznaczną identyfikację, ale nie czyni jej niemożliwą. Istnieją bowiem metody *fingerprintingu* urządzeń i przeglądarek pozwalające na identyfikację użytkowników pomiędzy przeglądarkami internetowymi zainstalowanymi na tym samym urządzeniu.

Gdyby podsumować metody identyfikacji użytkowników w dzisiejszym Internecie to ich niedługa lista prezentowałaby się następująco:

- Użycie *cookies* (w szczególności *third-party cookies*);
- Użycie *supercookies*;
- Użycie *fingerprintingu*.

Oczywiście nic nie stoi na przeszkodzie, aby każda z tych metod była używana w połączeniu z innymi.

2.3. Zagrożenia związane z fingerprintingiem

Al-Fannah i Mitchell [4, s. 1] konstatują, że identyfikacja za pomocą *fingerprintingu* jest znacznie trwalsza niż ta bazująca na *cookies*, praktycznie bez możliwości kontroli przez użytkowników i nietrywialna do wykrycia. Istnieją zatem przynajmniej trzy powody, dla których *fingerprinting* stanowi istotnie większe zagrożenie dla prywatności użytkowników (większe niż wszystkie dotychczasowe):

- Na tę chwilę nie są znane proste sposoby, aby z całkowitą pewnością wykryć, że dana platforma lub strona internetowa używa *fingerprintingu*;
- Użytkownicy mogą kontrolować moc śledzenia za pomocą *cookies*, regularnie usuwając je lub całkowicie blokując. Jak zostało nadmienione wcześniej, istnieją także regulacje prawne dotyczące użycia *cookies*. Nie ma jednak żadnych porównywalnych, równie prostych do użycia technik kontroli *fingerprintingu*;
- *Fingerprinting* (w przeciwieństwie do *cookies*) nie polega na jednej, konkretnej właściwości HTTP. *Fingerprinting* bazuje na wielu technologiach, aby zbierać różne informacje o właściwościach i konfiguracji przeglądarki lub systemu operacyjnego. Każda z tych informacji to szansa, aby zastosować *fingerprinting*.

Co więcej, *fingerprinting* może być użyty do tworzenia *supercookies*, czyli specjalnych *cookies*, które po usunięciu mogą zostać ponownie utworzone, jeśli ta sama przeglądarka zostanie wykryta przez użycie *fingerprintingu* [4, s. 2].

2.4. Możliwości ochrony przed fingerprintingiem

2.5. 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

Bibliografia

- [1] Edward E David Jr i Robert M Fano. „Some thoughts about the social implications of accessible computing”. W: *Proceedings of the November 30–December 1, 1965, fall joint computer conference, part I*. 1965, s. 243–247.
- [2] Peter Eckersley. „How unique is your web browser?” W: *International Symposium on Privacy Enhancing Technologies Symposium*. Springer. 2010, s. 1–18.
- [3] Kjeld Egevang, Paul Francis i in. *The IP network address translator (NAT)*. Spraw. tech. RFC 1631, may, 1994.
- [4] Nasser Mohammed Al-Fannah i Chris Mitchell. „Too little too late: can we control browser fingerprinting?” W: *Journal of Intellectual Capital* (2020).
- [5] Erik Flood i Joel Karlsson. „Browser fingerprinting”. W: (2012).
- [6] Vince Fuller i in. *Supernetting: An address assignment and aggregation strategy*. Spraw. tech. RFC-1338, June, 1992.
- [7] Robert Kahn i Vint Cerf. „A protocol for packet network intercommunication”. W: *IEEE Transactions on Communications* 22.5 (1974), s. 637–648.
- [8] Michal Kosinski, David Stillwell i Thore Graepel. „Private traits and attributes are predictable from digital records of human behavior”. W: *Proceedings of the national academy of sciences* 110.15 (2013), s. 5802–5805.
- [9] Eric Maigret. *Socjologia komunikacji i mediów*. Warszawa: Oficyna Naukowa, 2012.

- [10] M. G. Michael. *Ubervveillance and the social implications of microchip implants : emerging technologies*. Hershey, PA: Information Science Reference, 2014. ISBN: 978-1466645820.
- [11] Thomas Narten, Richard Draves i Suresh Krishnan. *Privacy extensions for stateless address autoconfiguration in IPv6*. Spraw. tech. RFC 3041, January, 2001.
- [12] Sheau Ng. „A brief history of entertainment technologies”. W: *Proceedings of the IEEE* 100.Special Centennial Issue (2012), s. 1386–1390.
- [13] Gil Press. „A very short history of the Internet and the Web”. W: *Forbes*. Luettavissa: [https://www. forbes. com/sites/gilpress/2015/01/02/a-very-short-history-of-the-internet-andthe-web-2](https://www.forbes.com/sites/gilpress/2015/01/02/a-very-short-history-of-the-internet-andthe-web-2) (2015).
- [14] Bruce Schneier. „The eternal value of privacy”. W: *Comment on Wired. com*, May (2006).
- [15] Edward Snowden. *Pamięć nieulotna*. Kraków: Insignis, 2019. ISBN: 978-83-66360-15-0.
- [16] Jun Wu. *The beauty of mathematics in computer science*. CRC Press, 2018.