Analiza i ocena bezpieczeństwa systemów usługowych i IoT
Ocena skuteczności różnych metod łamania haseł
Raport końcowy

Jan Pajdak
Wojciech Słowiński
Maria Filemonowicz

3 czerwca 2019

Prowadzący: Dr hab. inż. Grzegorz Kołaczek

Spis treści

1 Cel eksperymentu

Hasła tekstowe to obecnie najpopularniejsza metoda uwierzytelniana używana do ograniczania dostępu do zasobów takich jak serwisy internetowe czy konta pocztowe przez osoby nieupoważnione. Zabezpieczenia tego typu są łatwe w użyciu jednakże proste do złamania — w ramach eksperymentu analizowane będzie łamanie haseł przy użyciu algorytmów *BFM* oraz *Weira*.

Eksperyment będzie przeprowadzony przy użyciu bazy realnych haseł, które następnie będą badane pod kątem odporności na złamanie przez poszczególne algorytmy.

2 Plan eksperymentu

2.1 Źródło danych

Jako źródło danych wybrana została baza danych znaleziona w roku 2017 przez firmę z branży cyberbezpieczeństwa - *4iQ* [2]. Baza jest kompilacją informacji z 252 wycieków; zawiera loginy i hasła do ponad 1.4 miliarda kont. Całkowity rozmiar danych to 41.1 GB. Osoba odpowiedzialna za stworzenie bazy danych jest nieznana; dane zostały odkryte przez *4iQ* w *dark web* i można je obecnie pobrać przy użyciu sieci *torrent*.

Dane muszą zostać sformatowane przed użyciem ich w eksperymencie — są one porozdzielane na wiele plików oraz zawierają loginy i adresy poczty elektronicznej powiązane z kontami; te dodatkowe informacje są zbędne. Ze względu na ilość danych badany będzie podzbiór haseł.

2.2 Planowany przebieg eksperymentu

Algorytmy rozważane w niniejszym eksperymencie posiadają wysoką złożoność obliczeniową i czasową, przez co testowanie ich implementacji na dużych zbiorach haseł byłoby bardzo ograniczone przez moc obliczeniową dzisiejszych komputerów i czas potrzebny na przeprowadzenie takich badań. Aby umożliwić przeprowadzenie eksperymentu na o wiele większej bazie haseł, wykorzystane zostaną metody pozwalające na wyliczenie liczby prób potrzebnych do odgadnięcia hasła w obu algorytmach, nazywane *kalkulatorami* [3], dzięki którym wyznaczyć można siłę hasła.

2.3 Technologie

Do formatowania bazy haseł wykorzystany został *Python*.

Algorytmy oceniające skuteczność łamania haseł w poszczególnych algorytmach zostały zaimplementowane przy użyciu *Scala*.

2.4 Metoda oceny

2.4.1 BFM

Kalkulator oceny skuteczności algorytmu *BFM* działa następująco:

1. Na podstawie treningowego zbioru haseł określane są:

   Zbiór pojedynczych znaków alfabetu wraz z częstotliwością ich występowania jako pierwsza litera hasła

   Zbiór wszystkich możliwych digramów ułożonych ze znaków w alfabecie wraz z częstotliwością ich występowania, tworzony w następujący sposób:

   Zakładając zbiór treningowy złożony ze znaków *A, B, C*; Jeżeli znak A ma największe prawdopodobieństwo wystąpienia jako pierwszy, znak *B* ma największe prawdopodobieństwo wystąpienia po *A* a znak *C* ma największe prawdopodobieństwo wystąpienia po *B*, pierwszą próbą odgadnięcia hasła będzie *ABC*.

2. Na podstawie powstałych zbiorów określona zostaje kolejność zgadywania kolejnych znaków w haśle

3. Dla właściwego zbioru haseł wyznaczana jest liczba wymaganych prób potrzebnych do jego odgadnięcia:

   N - długość zbioru znaków w alfabecie

   L - długość zgadywanego hasła

   M - minimalna długość hasła

   CF - liczbę znaków sprawdzanych przed odgadnięciem właściwego pierwszego znaku hasła

   DF(i) - liczba digramów sprawdzanych przed odgadnięciem właściwego i-tego znaku hasła

   i - pozycja znaku w haśle

   k - k-ta próba odgadnięcia hasła

   X - liczba haseł krótszych niż zgadywane

   Y - liczba haseł z niepoprawnie odgadniętą pierwszą literą

   Z - liczba haseł z niepoprawnie zgadniętym i-tym znakiem

   Q - liczba prób potrzebnych do odgadnięcia zadanego hasła

   $X = \sum_{l=L}^{l=M} N^l$

   $Y = CF * N^{L-1}$

   $Z = \sum_{i=2}^{i=L} DF(i) * N^{L-(i+2)}$

   $Q = X + Y + Z$

   Jeżeli pierwszy znak nie zostanie odgadnięty poprawnie to wiemy, że algorytm podejmie N^{L-1} prób, zanim spróbuje odgadnąć hasło z innym znakiem na pierwszej pozycji

   Zgodnie z powyższym, dla k-tej próby odgadnięcia pierwszego znaku wiemy, że algorytm podejmie co najmniej $(k-1)N^{L-1}$ prób odgadnięcia hasła

2.4.2 Weir

Pojęcia używane w algorytmie Weir'a:

- struktura (*structure*) - opis określający typ każdego znaku w haśle

  L - litera

  D - cyfra

  S - znak specjalny

- (*terminal*) - instancja (hasło) spełniająca założenia struktury

- grupa prawdopodobieństwa (*probability group*) - grupa terminali o jednakowym prawdopodobieństwie wystąpienia w zbiorze treningowym

Przykład: struktura hasła "password123!@#" to "LLLLLLLLDDDSSS"

Proces tworzenia tablicy z której korzysta algorytm jest bardziej złożony niż w przypadku algorytmu BFM.

Algorytm 1: Pseudokod opisujący proces tworzenia tablicy używanej w kalkulatorze Weir'a

$T \leftarrow$ nowa tablica
forall *strukturys* **do**
  forall *grupy_prawdopodobiestwapg* **do**
    forall $lcs \in pg$ **do**
      c_i = liczba wystąpień lcs
      p_i = prawdopodobieństwo wystąpienia lcs w zbiorze treningowym
    end forall
    $probability = \prod p_i$
    $size = \prod c_i$
    $T.add: \quad pg, probability, size$
  end forall
end forall
$Sort(T)$ by $probability$

2.4.3 Estymator siły haseł Monte Carlo

W celu porównania ze sobą różnych algorytmów, służących do łamania haseł, skorzystano również z estymatora Monte Carlo [4], który szacuje liczbę prób potrzebnych do zgadnięcia danego hasła. Na podstawie ciągu treningowego wybierana jest próbka o zadanej długości n haseł na podstawie prawdopodobieństwa ich zgadnięcia w każdym z modeli ataków, wśród której estymator oszacowuje ile haseł z próbki zostanie złamane przed złamaniem podanego hasła, ustalając tym samym przybliżoną liczbę prób potrzebną na jego złamanie. Metoda ta nie jest dokładna, a wyniki są przybliżone i często zależą od różnych parametrów (takich jak wielkość próbki, wielkość ciągu treningowego), jednak pozwala na porównanie wielu modeli ataków oraz różnych polityk tworzenia haseł.

W implementacji Monte Carlo [1] wyróżnione zostały modele ataków bazujące na n-gramach, będące rozszerzeniem algorytmu BFM o dłuższe ciągi znaków, oraz model PCFG, który zaproponowany został przez Weira.

3 Przebieg eksperymentu

3.1 Przygotowanie danych

Dane znajdujące się w pobranej bazie są podzielone na 1 981 plików, w 54 folderach. Informacje w plikach są przechowywane w formacie *nazwa_użytkownika:hasło*. Rozmiar plików nie jest stały. Przed przystąpieniem do eksperymentu za pomocą skryptu napisanego w języku *Python* dane zostały uporządkowane tak, by utworzyć łatwy do analizy zbiór haseł. Algorytm wykonuje następujące operacje dla każdego pliku wchodzącego w skład pobranej bazy:

1. Wczytuje pojedynczą linię z pliku
2. Usuwa część linii zawierającą hasło
3. Przechodzi do następnej linii jeżeli hasło zawiera znaki inne niż:

   cyfry

   litery z alfabetu łacińskiego

   znaki specjalne

Tak przygotowana linia jest następnie przydzielana do kategorii i zapisana do odpowiedniego pliku. W eksperymencie rozróżniamy następujace kategorie haseł:

1. *raw* - każde hasło które nie zostało odrzucone po analizie znaków
2. *regular8* - hasło o długości min. 8 znaków
3. *regular16* - hasło o długości min. 16 znaków
4. *comprehensive8* - hasło o długości min. 8 znaków oraz zawierające co najmniej 1 dużą literę, 1 cyfrę i 1 znak specjalny
5. *comprehensive16* - hasło o długości min. 16 znaków oraz zawierające co najmniej 1 dużą literę, 1 cyfrę i 1 znak specjalny
6. *keepass16* - hasło o długości min. 16 znaków, które zostało losowo wygenerowane oraz zawierające co najmniej 1 dużą literę, 1 cyfrę i 1 znak specjalny

Wymogi kategorii *regular8* oraz *comprehensive8* bazują na popularnych wymogach dotyczących haseł stosowanych w różnych serwisach internetowych.

Zaledwie *0,5%* haseł wchodzących w skład bazy spełnia wymogi *comprehensive8*.

3.2 Ciąg treningowy i ciąg walidacyjny

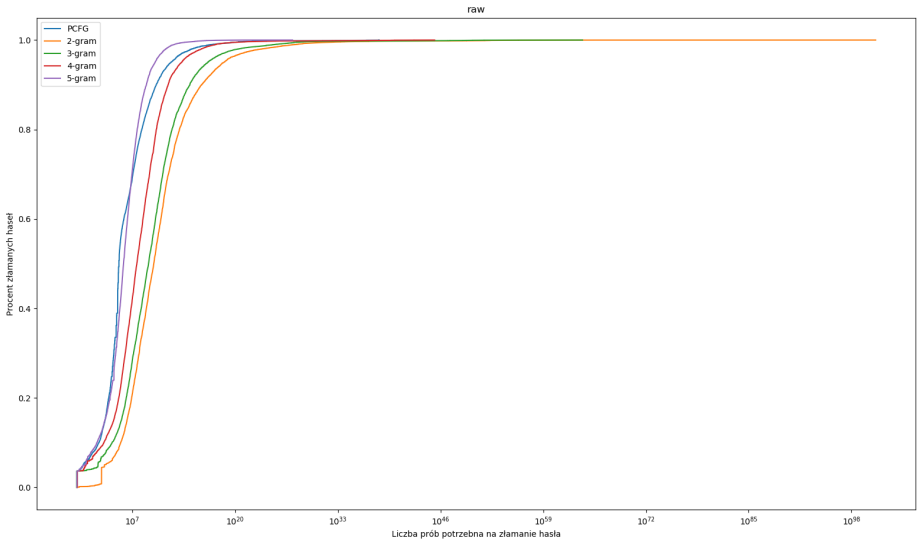
Ciąg treningowy to ważna część danych wejściowych, od których zależy skuteczność algorytmów. Dla każdej kategorii haseł został utworzony zbiór składający się z losowo wybranych haseł z odpowiadającej kategorii. Podobnie zostały stworzone ciągi walidacyjne.

Ciąg treningowy dla kategorii haseł *raw*, *comprehensive8*, *comprehensive16* i *regular8* składa się z 500 tysięcy haseł, natomiast ciąg walidacyjny z 50 tysięcy haseł. Dla polityki *regular8* zostały stworzone dwa ciągi treningowe - jeden o wielkości 500 tysięcy haseł i drugi o wielkości miliona haseł, aby sprawdzić wpływ wielkości ciągu treningowego na efektywność algorytmów. W poniższych badaniach wyróżnione zostaną w związku z tym kategorie nazwane *regular8 500k* oraz *regular8 mil*. W przypadku polityki *keepass16* również stworzone zostały dwa ciągi treningowe, oba o wielkości 500 tysięcy haseł, jednak aby sprawdzić jaki wpływ na efektywność algorytmów ma rodzaj haseł zawartych w ciągu treningowym, stworzonono zbiór składający się z haseł z polityki *keepass16* oraz zbiór składający się z haseł z polityki *regular16*, nazwyane w badaniach kolejno *keepass16 tsKeepass16* oraz *keepass16 tsRegular16*.

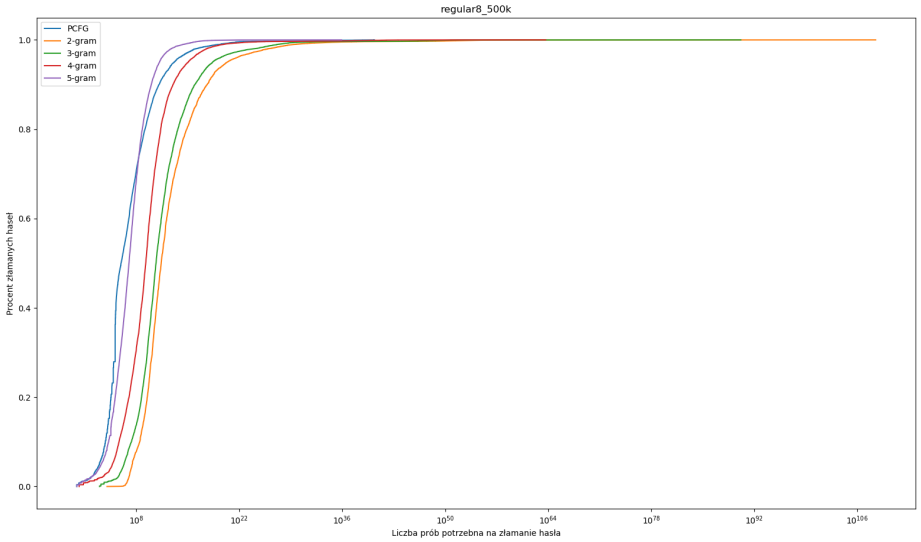
Dane służące jako zbiór treningowy kalkulatora oceny skuteczności algorytmu Weir'a zostały dodatkowo uporządkowane przy użyciu skryptu. Hasła zostały rozdzielone do plików odpowiadających ich strukturom. Dodatkowo, utworzony został plik zliczający wystąpienia każdego *lcs*.

4 Wyniki

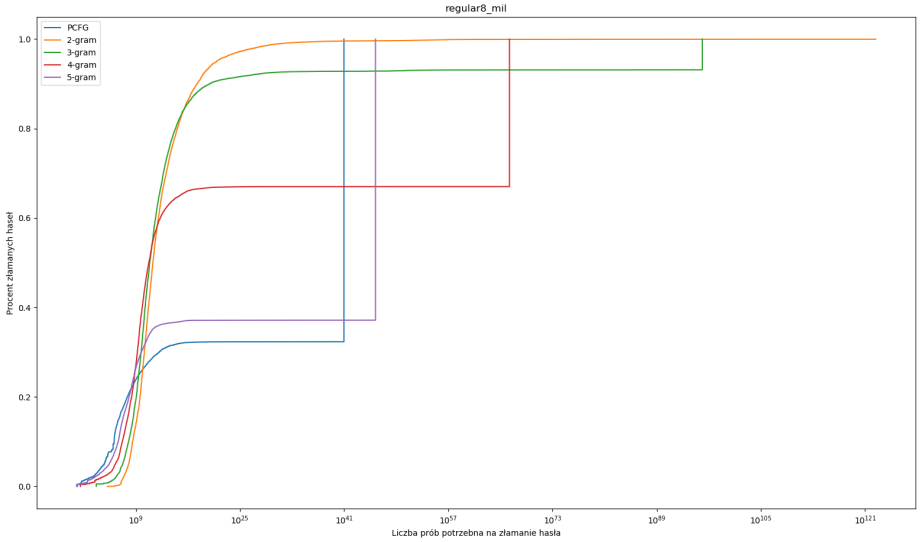
4.1 Porównanie wpływu polityki tworzenia haseł na ich siłę w kontekście różnych algorytmów



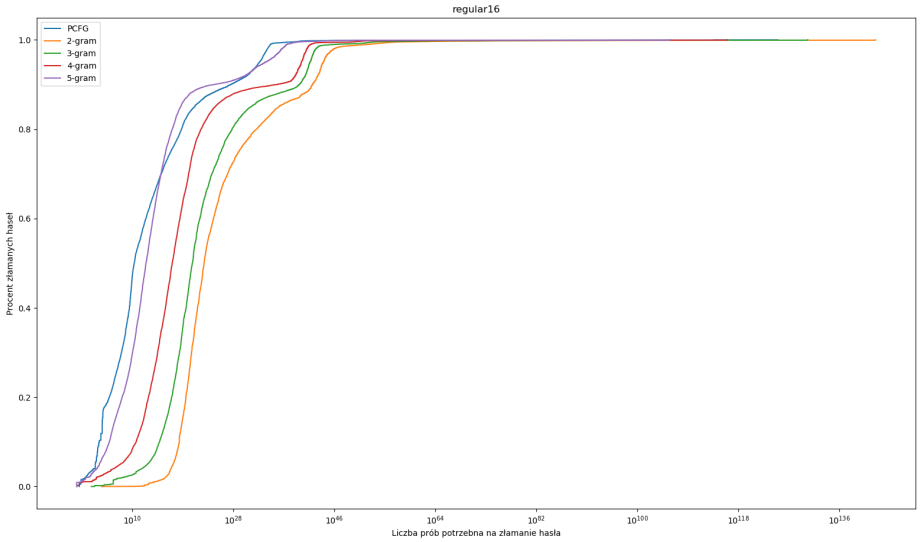
Rysunek 1: Wpływ rodzaju algorytmu na liczbę prób potrzebnych do złamania haseł nienależących do żadnej polityki



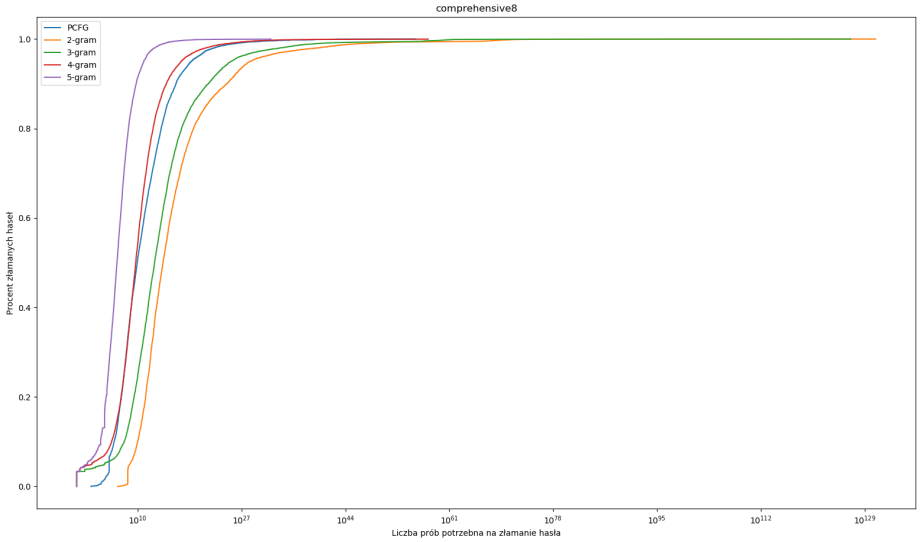
Rysunek 2: Wpływ rodzaju algorytmu na liczbę prób potrzebnych do złamania haseł w polityce regular8 przy ciągu treningowym wielkośći 500 tysięcy haseł



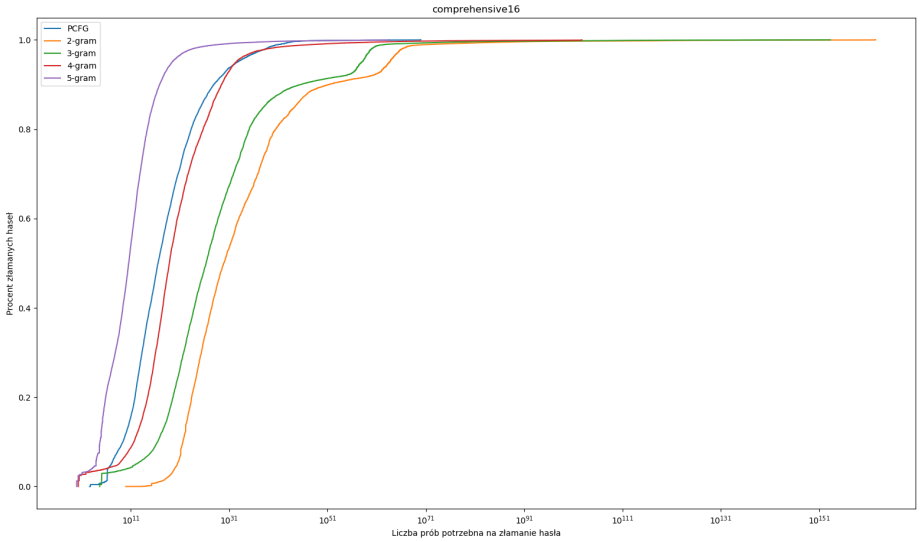
Rysunek 3: Wpływ rodzaju algorytmu na liczbę prób potrzebnych do złamania haseł w polityce regular8 przy ciągu treningowym wielkośći miliona haseł



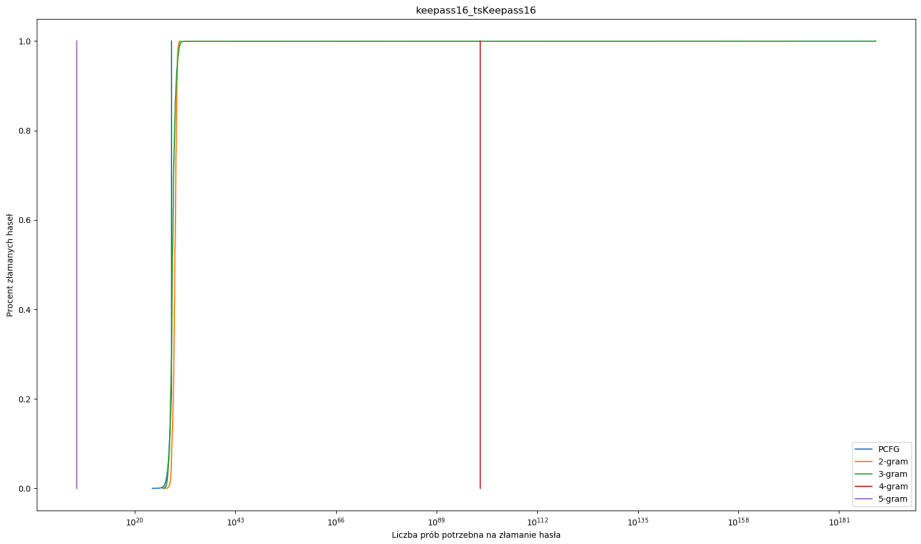
Rysunek 4: Wpływ rodzaju algorytmu na liczbę prób potrzebnych do złamania haseł w polityce regular16



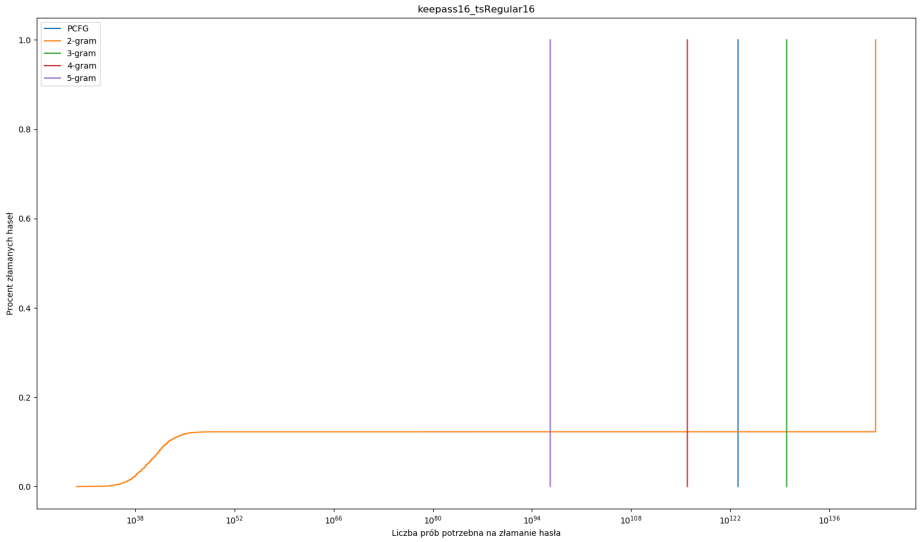
Rysunek 5: Wpływ rodzaju algorytmu na liczbę prób potrzebnych do złamania haseł w polityce comprehensive8



Rysunek 6: Wpływ rodzaju algorytmu na liczbę prób potrzebnych do złamania haseł w polityce comprehensive16



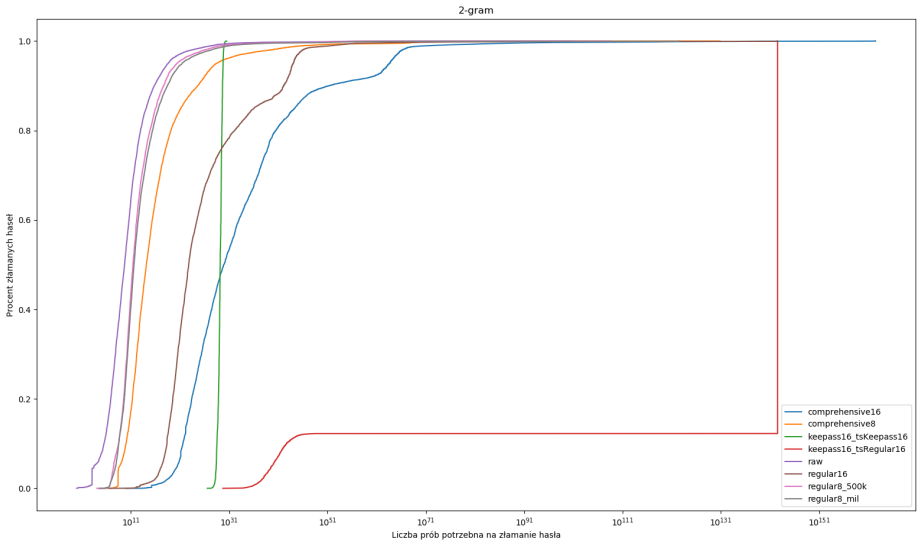
Rysunek 7: Wpływ rodzaju algorytmu na liczbę prób potrzebnych do złamania haseł w polityce keepass16 przy ciągu treningowym złożonym z haseł w polityce keepass16



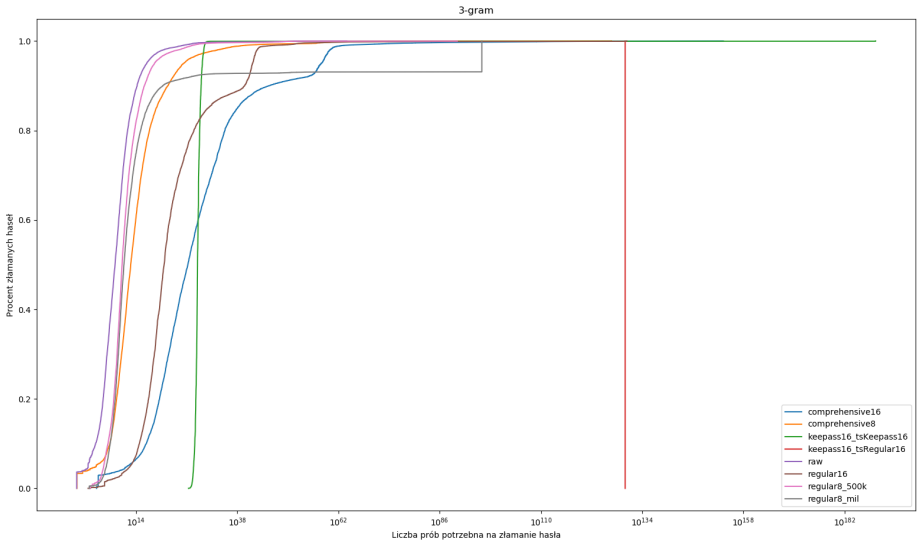
Rysunek 8: Wpływ rodzaju algorytmu na liczbę prób potrzebnych do złamania haseł w polityce keepass16 przy ciągu treningowym złożonym z haseł w polityce regular16

Rysunki 1 - 8 przedstawiają ile faktycznych prób musi wykonać każdy z algorytmów, aby odgadnąć procentową część wszystkich haseł z ciągu walidacyjnego.

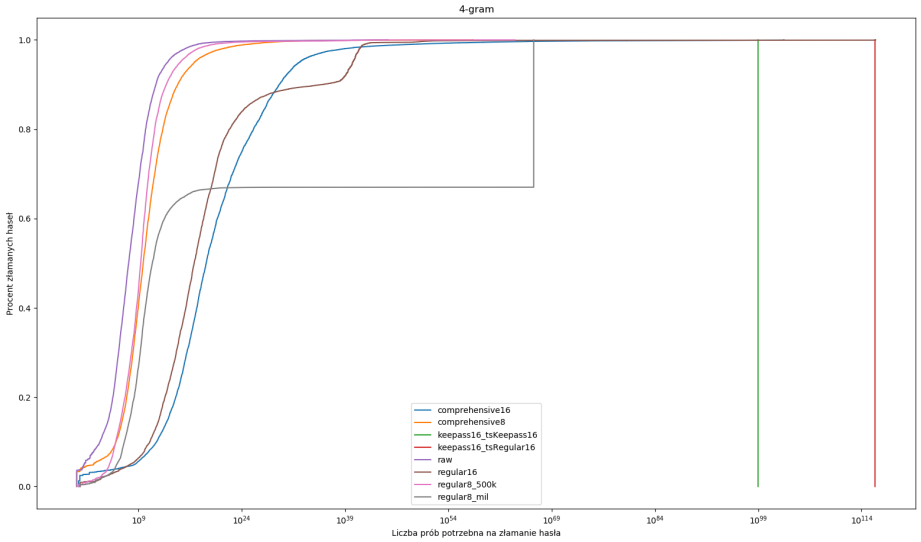
4.2 Porównanie wpływu rodzaju algorytmu na skuteczność łamania haseł w każdej z polityk



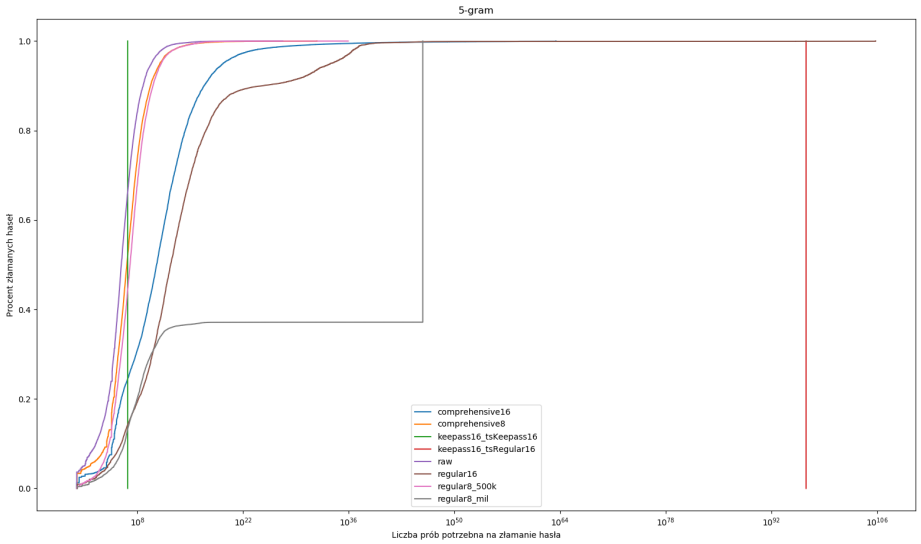
Rysunek 9: Skuteczność łamania haseł dla każdej polityki w algorytmie bazującym na 2-gramach (standardowy BFM)



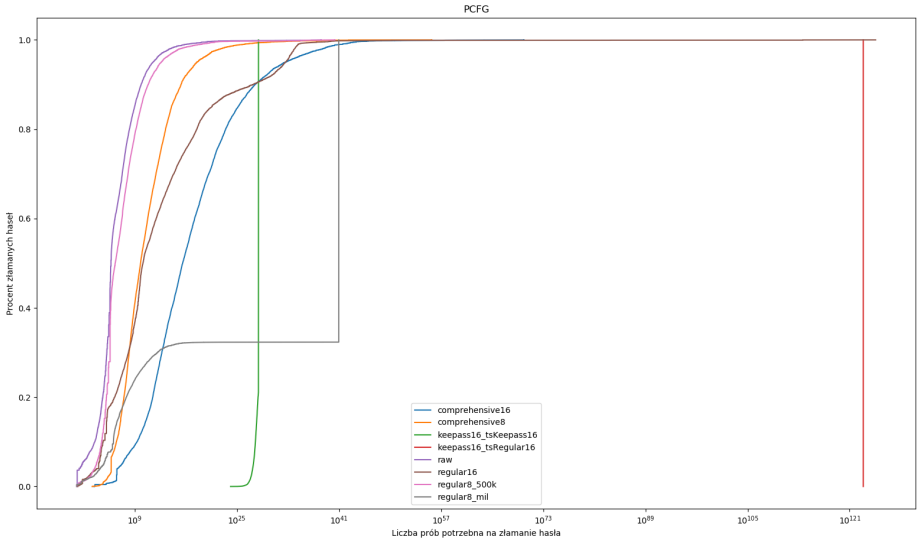
Rysunek 10: Skuteczność łamania haseł dla każdej polityki w algorytmie bazującym na 3-gramach



Rysunek 11: Skuteczność łamania haseł dla każdej polityki w algorytmie bazującym na 4-gramach



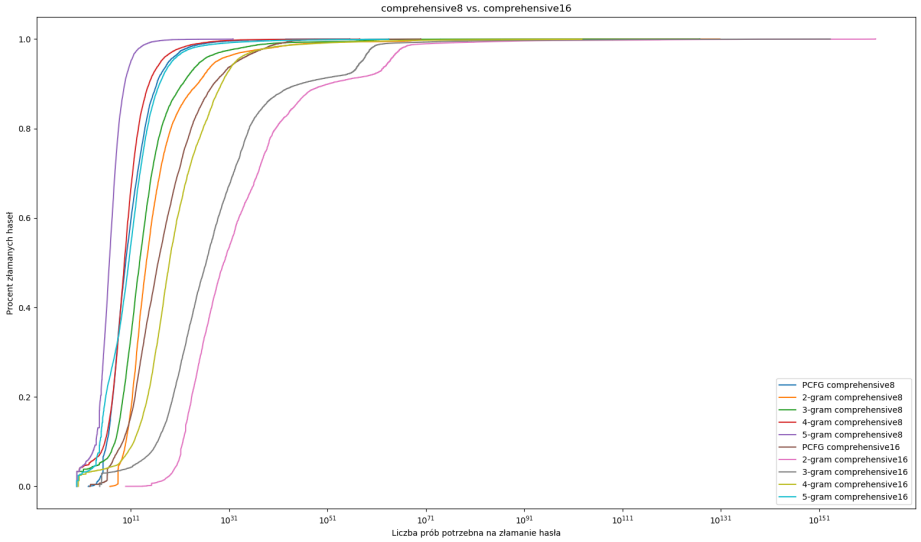
Rysunek 12: Skuteczność łamania haseł dla każdej polityki w algorytmie bazującym na 5-gramach



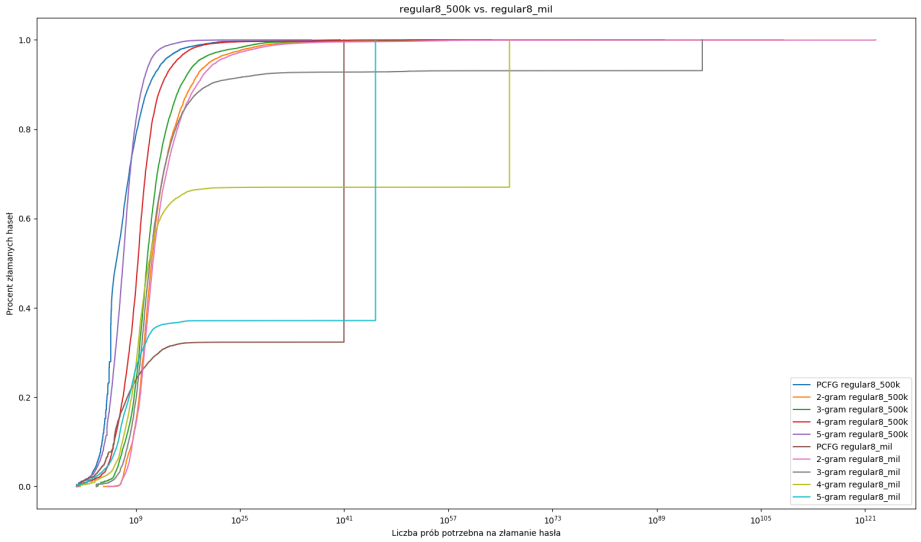
Rysunek 13: Skuteczność łamania haseł dla każdej polityki w algorytmie PCFG (algorytmie Weira)

Rysunki 9-13 przedstawiają liczbę prób potrzebną do złamania procentowej części wszystkich haseł z ciągu walidacyjnego.

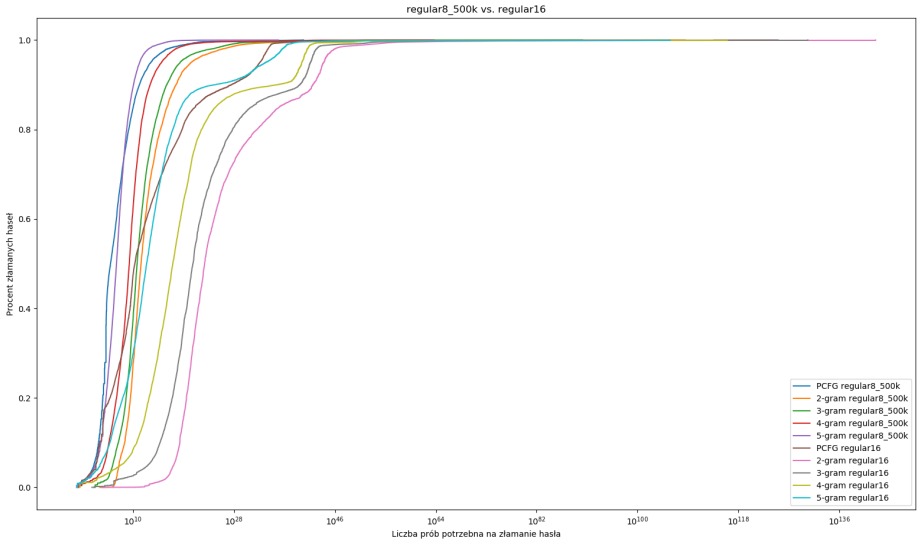
4.3 Porównanie poszczególnych polityk



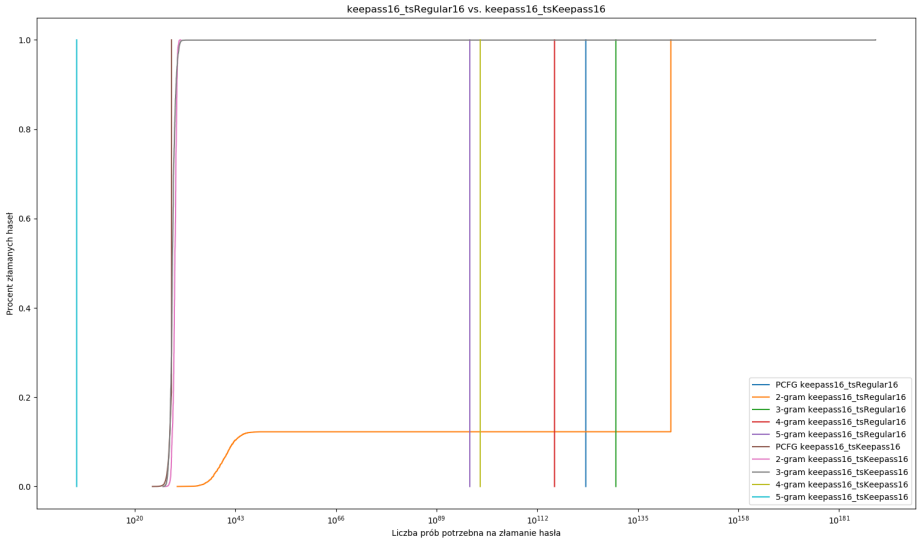
Rysunek 14: Porównanie polityki comprehensive8 i comprehensive16



Rysunek 15: Porównanie polityki regular8 na zbiorze treningowym wielkości 500 tysięcy haseł i wielkości miliona haseł



Rysunek 16: Porównanie polityki regular8 i regular16



Rysunek 17: Porównanie polityki keepass16 w przypadku zbioru treningowego zawierającego hasła z polityki regular16 i zbioru z hasłami z polity keepass16

5 Analiza wyników

5.1 Analiza wpływu polityki tworzenia haseł i rodzaju używanego algorytmu na skuteczność łamania haseł

Przeprowadzone badania potwierdzają zależność pomiędzy stopniem złożoności algorytmu a jego skutecznością niezależnie od wybranej polityki tworzenia haseł. Algorytm PCFG, który przeprowadza najbardziej szczegółową analizę haseł z ciągu treningowego, wyprzedza w większości polityk wszystkie algorytmy bazujące na n-gramach w liczbie prób, potrzebnych do złamania hasła. Podczas porównywania ze sobą n-gramów, również obserwowany jest wpływ długości ciągów znaków na skuteczność algorytmu - niezależnie od rodzaju łamanych haseł dłuższy ciąg znaków skutkował większą efektywnością. Bardzo zbliżoną skuteczność niezależnie od polityki haseł mają algorytmy PCFG i algorytm bazujący na 5-gramach. W przypadku polityk *comprehensive* i *keepass* algorytm oparty na 5-gramach wyprzedza algorytm PCFG.

Zarówno w przypadku algorytmów opierających się o 2-gramy, jak i 3-gramy, widoczna jest różnica wpływu polityki tworzenia haseł na ich siłę (w kontekście ich łamania przez w. w. algorytmy). Hasła, które nie miały żadnych zasad regulujących strukturę ich tworzenia, były bardziej podatne na złamanie niż w przypadku polityki *regular8* czy *comprehensive8*. Wśród tych dwóch polityk z kolei w około *90%* przypadków silniejsze były hasła, których struktura zawierała minimum jeden znak specjalny, jedną cyfrę i jedną wielką literę. Podobna tendencja widoczna jest również w przypadku haseł dwa razy dłuższych - *regular16* i *comprehensive16*.

W przypadku algorytmów opierających się na 2-gramach i 3-gramach siła haseł w zależności od regorystyczności polityki jest zachowana - najprościej jest złamać hasła bez żadnej polityki, a następnie coraz trudniej kolejno: *regular8*, *comprehensive8*, *regular16*, *comprehensive16* i najtrudniej *keepass16*. Podobne tendencje można zauważyć w przypadku algorytmów opierających się o 4-gramy, jednak hasła z polityki *keepass16* jest o wiele trudniej złamać - gdzie w przypadku algorytmu opierającego się o 5-gramy ich siła lekko spada. Algorytm PCFG zachowuje podobne prawidłowości w sile haseł w zależności od polityki.

5.2 Analiza wpływu wielkości zbioru treningowego na efektywność algorytmów

Na rysunku 15 została przedstawiona efektywność algorytmów łamiących hasła z polityki *regular8*, gdzie w jednym badaniu algorytmy miały do dyspozycji zbiór treningowy o wielkości 500 tysięcy haseł, a w drugim - miliona haseł. Dzięki nim można porównać wpływ wielkości zbioru treningowego na skuteczność algorytmów łamiących hasła. Z badań wynika, że wielkość zbioru treningowego faktycznie wpływa na liczbą prób potrzebnych do złamania hasła - dwukrotnie większy zbiór treningowy skutkuje dziesięciokrotnym zwiększeniem liczby prób potrzebnych do zgadnięcia hasła.

5.3 Analiza wpływu długości haseł na efektywność algorytmów

Aby porównać wpływ minimalnej długości hasła na ich siłę, przeprowadzone zostały badania porównujące ze sobą polityki *regular8* i *regular16* oraz *comprehensive8* i *comprehensive16*. W przypadku polityk *regular8* i *regular16* porównywane są wyniki przedstawione na wykresie 16, natomiast w politykach *comprehensive8* i *comprehensive16* na wykresie 14. Można dzięki nim zaobserwować, że efektywność algorytmów przy hasłach dwukrotnie dłuższych spada ponad stukrotnie - drastycznie wzrasta liczba prób potrzebnych do złamania hasła niezależnie od wykorzystanego algorytmu.

5.4 Analiza wpływu rodzaju haseł zawartych w ciągu treningowym na efektywność algorytmów

Na podstawie rysunku 17 porównywana jest siła haseł w polityce *keepass16* w przypadku uczenia algorytmów na dwóch różnych zbiorach treningowych - jednym opartym o hasła z tej polityki i drugim

opartym o hasła z polityki *regular16*. Zazwyczaj osoby, które mają na celu złamanie jakiegoś hasła, uwzględniają w zbiorze treningowym politykę tworzenia haseł w serwisie. W regułach tworzenia haseł w serwisie wymagania mogą być sprecyzowane co do długości hasła czy liczby znaków specjalnych, natomiast osoba łamiąca hasło najpewniej nie ma świadomości czy dane hasło jest słownikowe czy losowo generowane, więc najprawdopodobniej jako zbiór treningowy użyje zbioru haseł zawierających wszystkie rodzaje haseł spełniające wymagania serwisu.

Dzięki przeprowadzonym badaniom można zauważyć, że niezależnie od wybranego algorytmu zbiór uczący zawierający hasła z polityki, która odpowiada polityce łamanego hasła, czyli *keepass16* *tsKeepass16*, oferuje lepszą skuteczność łamania haseł niż zbiór treningowy składający się z haseł z innej polityki. Z punktu widzenia osoby łamiącej hasła najlepsze wyniki może osiągnąć znając jak najwięcej informacji o strukturze łamanego hasła - jego długości, zawartych w nim znaków czy jego znaczenia. Dzięki takim informacjom może lepiej dobrać zbiór treningowy i osiągnąć większą skuteczność w łamaniu haseł. Z punktu widzenia osoby tworzącej hasło lepiej jest używać haseł generowanych losowo, ponieważ - o ile osoba potencjalnie łamiąca hasło, nie ma świadomości, że hasło jest generowane losowo - takie hasła są wielokrotnie silniejsze w kontekście ich łamania.

6 Podsumowanie

Przeprowadzone badania potwierdziły słuszność najpopularniejszych polityk tworzenia haseł uznawanych za skuteczne. Hasła dłuższe i zawierające znaki specjalne są trudniejsze do złamania, choć najbezpieczniejszą metodą tworzenia haseł są jest ich losowe generowanie. Niezależnie od wykorzystanego algorytmu łamiącego hasła są one nieporównywalnie silniejsze, a większość osób łamiących hasła dysponuje ciągiem treningowym w postaci zbioru zawierającego hasła tworzone ze słów. W analizowanej bazie haseł można jednak znaleźć ogromną liczbę haseł, które są bardzo proste i przewidywalne, co wskazuje na niedbałość lub nieświadomość osób zabezpieczających swoje konta.

Efektywność badanych algorytmów do łamania haseł rośnie wraz ze stopniem ich złożoności. Czym więcej operacji analizujących ciąg treningowy musi zostać wykonanych podczas pracy algorytmu, tym mniej prób potrzeba, aby hasło zostało złamane. Na skuteczność algorytmu ma duży wpływ wielkość ciągu treningowego oraz odpowiednia kategoria haseł w nim zawarta - szybciej złamane zostaną hasła, które odpowiadają polityką tym zawartym w ciągu treningowym. Osoba łamiąca hasła może zyskać największą przewagę znając jak najwięcej informacji o strukturze łamanego hasła (jego długości, zawieranych znaków, losowości lub znaczenia), ponieważ dzięki temu może dobrze dopasować ciąg treningowy, a tym samym uzyskać najlepsze rezultaty.

Literatura

[1] *Monte Carlo password checking implementation*. URL `https://github.com/matteodellamico/montecarlopwd`. Dostęp 10.05.2018.

[2] Julio Casal. *1.4 Billion Clear Text Credentials Discovered in a Single Database*, 2017. URL `medium.com/4iqdelvedeep/1-4-billion-clear-text-credentials-discovered-in-a-single-database-3131d0a1ae14`. Dostęp 08.04.2018.

[3] Patrick Gage Kelley Saranga Komanduri Michelle L. Mazurek Richard Shay Timothy Vidas Lujo Bauer Nicolas Christin Lorrie Faith Cranor and Julio Lopez. *Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms*, 2012. URL `archive.ece.cmu.edu/~lbauer/papers/2012/oakland2012-guessing.pdf`. Dostęp 08.04.2018.

[4] Maurizio Filippone Matteo Dell'Amico. *Monte Carlo Strength Evaluation: Fast and Reliable Password Checking*, 2015. URL `http://www.eurecom.fr/~filippon/Publications/ccs15.pdf`. Dostęp 10.05.2018.