Gabriel Wechta, 250111

02.11.2020

1 Omówienie problemu

Treść zadania znajduje się na stronie doktora Zagórskiego. Scenariusz jest następujący: przechwyciliśmy kilkanaście szyfrogramów, o których wiemy, że były szyfrowane za pomocą szyfra strumieniowego, o kórym z kolei nic nie wiemy, przy wykorzystaniu tego samego klucza i tego samego i niezmienniczego generatora do generowania bitów pseudolosowych, co jest oczywiście ogromnym nadużyciem bezpieczeństwa szyfru strumieniowego.
Bez skrupułów wykorzystamy to nadużycie.

2 Wybór techniki

Metodą zasugerowaną w książce "Cryptography. Theory and practice" - Douglas R. Stinson jest analiza i wykorzystanie częstotliwości występowania liter w języku naturalnym. Ta metoda ma niestety życzeniowe założenia dotyczące tesktu, między innymi, że znany jest język w jakim pisane są wiadomości, oraz że znamy wielkość liter. Treść zadania nie podaje nam żadnych informacji na ten temat. Ponadto powyższa metoda wprowadza pewne prawdopodobieństwo poprawności rozwiązania.
Zdecydowałem się więc na wykorzytsanie metody szukania spacji w tekście, dającej pewne wyniki. Przy założeniu, że używamy tylko znaków z klawiatury (ASCII), metoda działa niezależnie od kodowania (ASCII/UTF-8/ISO-8859-2), ze względu na to, że wartości przypisane literą i spacji są takie same we wszytskich kodowaniach. Ponadto była to jedyna metoda omówiona na wykładzie wykorzystywana do łamania szyfrów strumieniowych.

3 Technika

Niech:

1. m_i – wiadomość, plaintext.
2. k – klucz.
3. $c_i = m_i \oplus k$ – zaszyfrowana wiadomość.

Fakty dotyczące XOR:

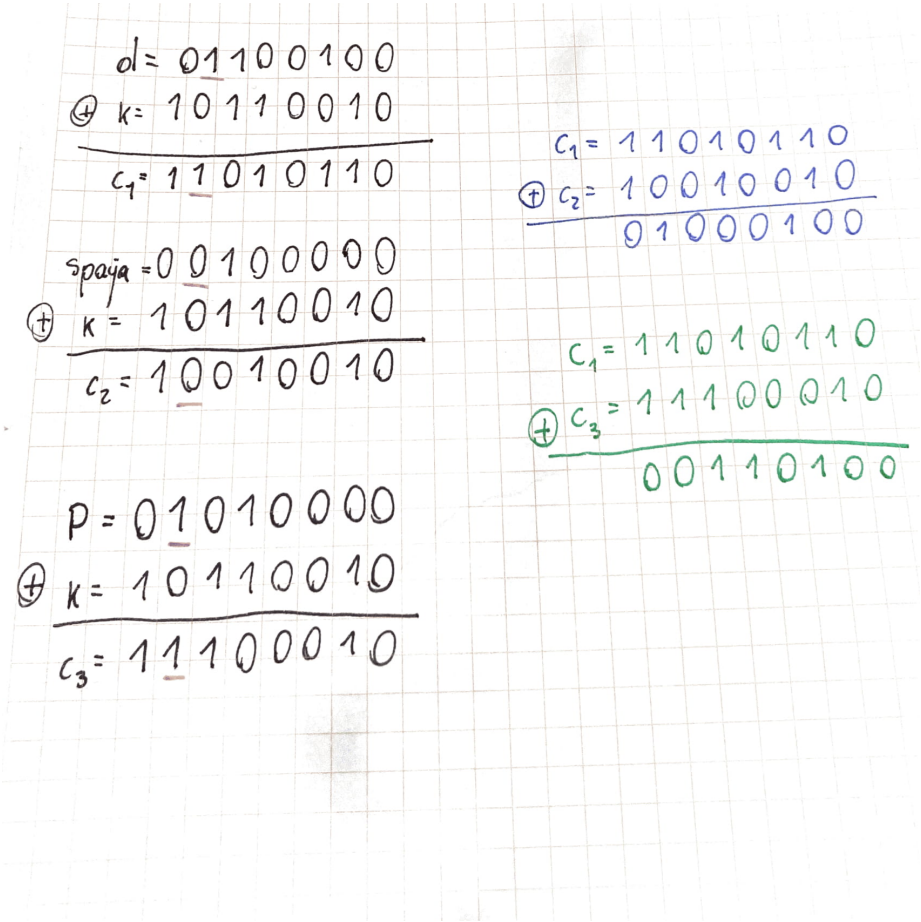
1. $c_i \oplus c_j = (m_i \oplus k) \oplus (m_j \oplus k) = m_i \oplus m_j$.
2. Dla znaków ASCII z [a-zA-Z] : $(c_i \oplus c_j)_{10} \geqslant 65 \Leftrightarrow c_i = 32 \vee c_j = 32$
3. $k = m_i \oplus c_i$

Drugi fakt wynika z tego, że spacja jest jedynym znakiem którego 7 bit jest wyzerowany i występuje w tekście. Wykorzystując fakt trzeci i znając wartość m_i (znamy gdy jest równa spacji), możemy obliczyć klucz. Znając klucz i wiedząc, że był wykorzystany do zakodowania każdego znaku w odpowiadającej mu kolumnie, możemy odkodować wszystkie znaki w tej kolumnie.

4 Algorytm

Nie będę opisywał algorytmu pseudokodem, ponieważ nie widzę w tym celu, opowiem jak działa. Zapisujemy zakodowane wiadomości bajtami. Bierzemy pierwszy bajt z pierwszej zakodowanej wiadomości porównujemy go z każdym bajtem w tej kolumnie, jeśli dla każego XOR'a zwróci wartość wieskzą niż 64, mamy pewność, że to spacja. Wyjaśnioną wyżej metodą odkodowujemy resztę znaków w kolumnie.

5 Rysunek agitacyjny



Powyższy rysuenk pokazuje poprawność idei. Spacja jest jedynym znakiem występującym w tekście, którego XOR z każdym innym znakiem w tej samej kolumnie, zwróci liczbę, której siódmy bit od prawej będzie inny od pozostałych. Co ozancza, że XOR c_1 i c_2 zwróci liczbę większą od 64.

6 Ciąg dalszy techniki

Nietrudno zauważyć, że odkodujemy wyłącznie te kolumny, w których występują spacje. Może nie brzmi to jak perfect case ale prawda jest taka, że jest to dobre. Odkodowane znaki są pewne, nie musimy się martwić o częstotliwość wystąpień liter w tekście, na dobrą sprawę to nawet nie musimy wiedzieć w jakim języku wysyłane są te wiadomości.
Ale co dalej?
Output programu dla danych generowanych przez indeks 123456.

```
gabriel$> ? a c z ? n s k ?  w y ? a s ? i    ? ?
s k a d  w z ? e ? ? y  ? ? ?
p r o t ? ? t ? ?  Z e b y  ? ? ? n i ?
b ? ? ? n ? c ? ? i  t e r e ? ? w
S l ? w o m i r a ?  N ? ? ? k a ?
```

Odgadnięcie brakujących liter w tym wyniku nie powinno stanowić problemu dla każdego kto nie obudził się wczoraj ze snu zimowego. Teraz technika przechodzi w technikę krzyżówkową. Odgadujemy literę, liczymy XOR z odpowiadającej jej bajtem kodu i dostajemy klucz. Jeżeli nie jesteśmy w stanie odgadnąc litery, możemy wziąć pod lupę dowolne inne zdanie.

7 Testy

Dane do testów są dołączone do kodu. Zamierzam o nich opowiedzieć pokazując wyniki.

7.1 Długość szyfrogramu

W zaprezentowanej przeze mnie metodzie, szyfrogramy dłuższe dają nam większą szansę na znalezienie spacji. Jeżeli zależy nam na odkodowaniu wszystkich wiadomości to ma to znaczenie, jeżeli wyłącznie na odkodowaniu ostatniego szyfrogramu to szyfrogramy dłuższe niż on są w zasadzie bez znaczenia.

7.2 Liczba szyfrogramów

Oprócz większej szansy na znalezienie spacji, większa liczba szyfrogramów pozwala nam też mieć więcej zdań, do których możemy sie odwołać przy przejściu na ręczne szukanie. Dla długich zdań usunięcie połowy szyfrogramów dalej pozwala na przeczytanie zagadki.

7.3 Typ szyfru

Nie udało mi się znaleźć online toola do kodowanaia wiadmości przy użcyiu wymienionyych szyfrów, niemniej wydaje mi się, że tak długo jak każda wiadomość jest kodowana tym samym kluczem (generatorem klucza) oraz n-temu blokowi wiadomości m_i odpowiada n-ty blok kodu oraz n-temu blokowi wiadomości m_j odpowiada ten sam blok kodu oraz szyfrogram uzyskiwany jest XORem i pozostajemy w jednym typie kodowania znaków, typ szyfrowania jest bez znaczenia.

7.4 Kodowania znaków

Dochodzimy ponownie do sytuacji, w której musimy założyć pewne rzeczy o przesyłanych wiadomościach, jeżeli zawierają wyłącznie znaki ze zbioru ASCII, ich kodowanie nie ma znaczenia dla algorytmu ponieważ ich zapis bajtowy nie różni się od zapisu w ASCII. Spacja ma wartosć 32 - decymalnie w ASCII/UTF-8/ISO-8859-2, 'a' ma 97 itd.
Natomiast jeżeli w wiadomościach pojawią się znaki z poza ASCII jak na przykład '¢', czy '€' to mamy problem, algorytm nie zadziała, ze względu na to, że na zapis znaków z poza ASCII potrzebne są różne liczby bajtów co zupełnie łamie core algorytmu.