Hello, World!

Вижнерова шифра

Миодраг Живковић, Београд

1. Шифровање, дешифровање, криптоанализа

Алиса жели да пошаље поруку Бобану, али тако да избегне да Цица, која може да пресретне ту поруку, буде у стању да је прочита. Због тога се Алиса и Бобан договарају који ће **алгоритам** и **кључ** користити за шифровање, односно дешифровање поруке. Пре него што пошаље поруку, Алиса је **шифрује**. Кад Бобан прими поруку, да би могао да је прочита, он је најпре **дешифрује**. Цица пресреће поруку, али не може да је дешифрује као Бобан, јер **зна** који су Алиса и Бобан алгоритам шифровања изабрали, али **не зна** кључ. Она може да спроведе **криптоанализу** добијеног **шифрата**, тј. да полазећи од шифрата некако реконструише поруку, иако не зна кључ. За детаље видети [3].

Тешко је преценити значај **криптографије**, науке о поступцима шифровања. У току другог светског рата Немци су за тајно преношење најповерљивијих података, поред осталог, користили уређај Енигма. Пољаци су и пре рата успели да реконструишу одговарајући поступак шифровања. Енглези су преузели њихове резултате и наставили са нападом на шифру. У ту сврху развили су рачунар Колос (Colossus, [1]), који због тајности није спомињан у историји развоја рачунара до 1990. године. Најмоћнији рачунари су се одувек користили у војне сврхе, пре свега за моделирање нуклеарног оружја, разбијање шифри и моделирање времена [6]. Књига Дејвида Кана [4] је једна од најинтересантнијих књига о историји криптографије.

2. Шифра транслације

Један од најједноставнијих поступака шифровања је **шифра транслације**. Претпоставимо да се азбука којом је написана порука састоји од n слова; за нашу ћирилицу је $n = 30$ (видети табелу 1), а за енглеску латиницу је $n = 26$. Словима азбуке одговарају редни бројеви од 0 до $n - 1$. Кључ k је било који број од 0 до $n - 1$. Ако се порука састоји од N слова, чији су редни бројеви у азбуци редом p_1, p_2, … , p_N онда се редни бројеви c_1, c_2, … , c_N слова шифрата добијају као остаци при дељењу са n после сабирања се k, тј. за $i = 1, 2, \dots, N$ је $c_i = (p_i + k) \bmod n$. Овде је mod ознака за операцију рачунања остатка при дељењу првог аргумента другим. Дешифровање је инверзна операција: i-то слово поруке, p_i, добија се као остатак при дељењу са n разлике $k - c_i$, тј. $p_i = (k - c_i) \bmod n$.

Слово	А	Б	В	Г	Д	Ђ	Е	Ж	З	И	Ј	К	Л	Љ	М
Индекс	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Слово	Н	Њ	О	П	Р	С	Т	Ћ	У	Ф	Х	Ц	Ч	Џ	Ш
Индекс	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29

Табела 1. Редни бројеви ћириличних слова.

Шифровање и дешифровање могу се уместо као операције са редним бројевима слова схватити као операције са самим словима. На пример, шифровање "А" померањем за 7 места може се написати као сабирање слова "А" са (седмим) словом "Ж" : "А" + "Ж" = "Ж", а дешифровање "Ж" истим кључем "Ж" даје "Ж" − "Ж"="А".

Пример 1. Алиса жели да Бобану пошаље поруку „АНА ВОЛИ МИЛОВАНА“. Ана и Бобан договорили су се да користе шифру транслације са кључем "Ж", седмим словом у азбуци. Приликом шифровања свако слово поруке замењује се словом које се налази 7 места иза њега у азбуци, видети табелу 2. Користећи ову табелу и замењујући једно по једно слово поруке, Алиса добија шифрат "ЖЋЖ ИФРЊ ТЊРФИЖЋЖ". Бобан добијени шифрат дешифрује користећи исту табелу у обрнутом смеру, односно замењујући свако слово словом које је 7 слова испред њега у азбуци. Наравно, он добија поруку „АНА ВОЛИ МИЛОВАНА“.

Порука	А	Б	В	Г	Д	Ђ	Е	Ж	З	И	Ј	К	Л	Љ	М
Шифрат	Ж	З	И	Ј	К	Л	Љ	М	Н	Њ	О	П	Р	С	Т
Порука	Н	Њ	О	П	Р	С	Т	Ћ	У	Ф	Х	Ц	Ч	Џ	Ш
Шифрат	Ћ	У	Ф	Х	Ц	Ч	Џ	Ш	А	Б	В	Г	Д	Ђ	Е

Табела 2. Шифра транслације за 7 места дуж азбуке, односно сабирање са словом "Ж" као кључем.

Позната Цезарова шифра је уствари шифра транслације са фиксираним кључем 3. Другим речима, Цезарова шифра се од других разликује по томе што уствари нема кључ: непромењена порука увек даје непромењени шифрат. Други познати пример шифре транслације дискретно се појављује у филму Одисеја у свемиру 2001: главни „јунак“ је рачунар HAL, чије се име добија тако што се у речи IBM свако слово замени претходним словом латинице.

Како Цица може да **нападне** овај систем, односно да одреди поруку ако зна да је искоришћена шифра транслације, а не зна кључ? Пошто могућих кључева у овом систему има само 30, она може да дешифрује шифрат свим кључевима, видети табелу 3. Пошто се само за кључ 7 дешифровањем

добија смислена порука, Цица на тај начин успева да **потпуном претрагом** одреди кључ и прочита поруку.

.

Кључ	Резултат дешифровања поруке "ЖЋЖ ИФРЊ ТЊРФИЖЋЖ"														
А	Ж	Ћ	Ж	И	Ф	Р	Њ	Т	Њ	Р	Ф	И	Ж	Ћ	Ж
Б	Е	Т	Е	З	У	П	Н	С	Н	П	У	З	Е	Т	Е
В	Ђ	С	Ђ	Ж	Ћ	О	М	Р	М	О	Ћ	Ж	Ђ	С	Ђ
Г	Д	Р	Д	Е	Т	Њ	Љ	П	Љ	Њ	Т	Е	Д	Р	Д
Д	Г	П	Г	Ђ	С	Н	Л	О	Л	Н	С	Ђ	Г	П	Г
Ђ	В	О	В	Д	Р	М	К	Њ	К	М	Р	Д	В	О	В
Е	Б	Њ	Б	Г	П	Љ	Ј	Н	Ј	Љ	П	Г	Б	Њ	Б
Ж	А	Н	А	В	О	Л	И	М	И	Л	О	В	А	Н	А
З	Ш	М	Ш	Б	Њ	К	З	Љ	З	К	Њ	Б	Ш	М	Ш
...	...														
Ц	И	Ф	И	К	Ц	Т	П	У	П	Т	Ц	К	И	Ф	И
Ш	З	У	З	Ј	Х	С	О	Ћ	О	С	Х	Ј	З	У	З

Табела 3. Разбијање шифре транслације потпуном претрагом.

3. Полиалфабетске шифре

Као што смо видели, шифра транслације је уствари сабирање свих слова поруке једним истим словом кључа. Због тога се каже да је шифра транслације моноалфабетска шифра. Природно уопштење моноалфабетске шифре су **полиалфабетске шифре**, код којих се за шифровање сваког наредног слова поруке $p_1, p_2, \dots, p_N$ користи ново слово **низа кључа** $k_1, k_2, \dots, k_N : c_i = (p_i + k_i) \bmod n, i = 1, 2, \dots, N$. Дешифровање је инверзна операција којом се од шифрата поново добија порука: $p_i = (k_i - c_i) \bmod n$.

Идеја да се несигурне моноалфабетске шифре замене полиалфабетским доста је стара. Италијан Леон Алберти (Leon Battista Alberti, 15. век) смислио је специјални диск за шифровање, што је поједноставило сабирање слова поруке са словом кључа. Други Италијан, Ђован Беласо (Giovan Battista Bellaso) је 1466. године описао полиалфабетску шифру у којој су се алфабети повремено мењали, што је назначено у шифрату; сам низ алфабета (транслација) договара се пре слања поруке. Упркос томе, проналазак такве шифре погрешно се приписује Француз Блезу Вижнеру (Blaise de Vigenere, 16. век), дипломати, криптографу, преводиоцу и алхемичару. Све до 19. века рачунало се да се ова шифра не може разбити. Први општи метод напада на шифру објавио је 1863. године пруски официр Фридрих Касиски (Friedrich Kasiski).

С обзиром на то како се формира низ кључа, постоји неколико типова полиалфабетских шифри:

- Као низ кључа може се користити неки други (унапред договорени) текст. У претходним вековима су шпијуни за шифровање својих порука често користили Библију, рачунајући да то не може да изазове подозрење. При томе су уз поруку могли да наведу рецимо страну Библије од које почиње низ кључа.
- **Случајна шифра** је полиалфабетска шифра код које се као низ кључа користи изгенерисани низ случајних независно изабраних слова исте дужине као и порука. Ова шифра се зове и **апсолутно тајна шифра.** Ова шифра се не може разбити, јер су све могуће (и смислене) поруке исте дужине једнако вероватне, па нема начина да се установи која је од њих права. Једном искоришћени низ кључа не сме се никада поново искористити. Уколико се две различите поруке шифрују истим низом кључа, онда се обе могу одредити тзв. нападом Казиског.
- **Вижнерова шифра** као низ кључа користи довољно пута поновљену једну исту реч - **кључ**. На пример, ако је кључ реч "ПЕТАР", онда је низ кључа текст "ПЕТАРПЕТАРПЕТАР ... ".

Пример 2. Као пример поруке искористићемо почетак Андрићевог романа „На Дрини ћуприја“:

„Већим делом свога тока река Дрина протиче кроз тесне гудуре између стрмих планина или кроз дубоке кањоне окомито одсечених обала. Само на неколико места речног тока њене се обале проширују у отворене долине и стварају, било на једној било на обе стране реке, жупне, делимично равне, делимично таласасте пределе, подесне за обрађивање и насеља. Такво једно проширење настаје и овде, код Вишеграда, на месту где Дрина избија у наглом завоју из дубоког и уског теснаца који стварају Буткове Стијене и Узавничке планине. Заокрет који ту прави Дрина необично је оштар а планине са обе стране тако су стрме и толико ублизу да изгледају као затворен масив из којег река извире право, као из мрког зида. Али ту се планине одједном размичу у неправилан амфитеатар чији промер на најширем месту није већи од петнаестак километара ваздушне линије.“

Добијени шифрат искористићемо за демонстрацију поступка разбијања шифре у наредна два примера. Примери су направљени програмом из мастер рада [2]. Пре шифровања у тексту се остављају само слова, тако што се сва слова претварају у мала:

„већимделомсвогатокарекадринапротичекрозтеснегудуреизмеђустрмихпла
нинаиликроздубокекањонеокомитоодсеченихобаласамонанеколикоместаре
чногтокањенесеобалепроширујууотворенедолинеиствarајубилонаједнојбил
онаобестранерекежупнеделимичноравнеделимичноталасастепределеподес
незаобрађивањеинасељатаквоједнопроширењенастајеиовдекодвишеградан
аместугдедринаизбијаунагломзавојуиздубокогиускогтеснацакојиствarајубу
тковестијенеиузавничкепланинезаокреткојитуправидринанеобичнојеоштар
апланинесаобестранетакосустрмеитоликоублизудаизгледајукаозатворенмас
ивизкојегрекаизвиреправокаоизмркогзидаалитусепланинеодједномразмичу
унеправиланамфитеатарчијипромернанајширемместунијевећиодпетнаестак
километараваздушнелиније“

Шифровањем кључем "петар" добија се шифрат:

"слљигћлгогзззгриувазфотдзчттпзђчачхшхззјфцеећкјмрхчмђефкцлргчбилрг
неаџанврецјмбешлвађђтчоаћсатеђјкењфтахерегаипсзнрглвобчозмхзчтрхнтз
гјђотњхглкеерегежжусизкњмуеиззрхглхобчтчииизтррџшћибђттјхћтзјсчпзн
рђжчсјжееезфочжлетчдханђињгујатглхебчсачдђчтлрзектхехчдхалиоуфцееч
пуђррунуађфнеаифрттршззјхћтзпзђђархдлеаииебеџђзхеађјуипфијаупттмхз
чмгуфјјидпншбџцемнртпзмчпззјлчмхусћозгџкцвоћилкнрмевошчцлвржебус
кчвотфцлишфтчилцеунџночпбптанхцезкзфчвошччмпзпзадзчттнхђжачдђњч
опиејажаееидфцтосфцлрргллаађцмсјжсчијђпакекжгичкјтичтпчдрџшваецел
вежлемрзнуичшубеђжлваџцзархехтвешезичвхвоћцнхаранлуиффгадчтчоуџл
хневхтзгчгмудффјатчптнрваатхпчтрњчњапзђсчрдпттјпчхчмгфцлудчњчвхјн
зджфчеахзчткачпзмхиејатпмхупглгидчњч"

4. Разбијање Вижнерове шифре

Разбијање Вижнерове шифре састоји се од две фазе: одређивање дужине кључа, а затим одређивање самог кључа. Да бисмо могли да разумемо овај напад, упознаћемо се најпре са појмом индекса коинциденције који је увео амерички криптограф Вилијем Фридман [5].

Индекс коинциденције

Напад на Вижнерову шифру заснива се на примени тзв. индекса коинциденције. Индекс коинциденције текста једнак је вероватноћи да два случајно независно изабрана слова из текста буду једнака.

Нека је $P = p_1, p_2, \dots, p_N$ ћирилични текст од n слова и нека је n_i број појављивања i-тог слова азбуке, $i = 0,1, \dots, n-1$ (за ћирилицу је $n = 30$). Тада је индекс коициденције $IC(P)$ текста P једнак збиру вероватноћа да та два изабрана слова буду оба једнака "А",...,"Ж", односно

$$IC(P) = \sum_{i=0}^{n-1} \frac{\frac{n_i(n_i-1)}{2}}{\frac{N(N-1)}{2}} = \sum_{i=0}^{n-1} \frac{n_i(n_i-1)}{N(N-1)}.$$

Пример 2. У поруци "АНА ВОЛИ МИЛОВАНА" слова А, В, И, Л, М, Н, О појављују се редом $4, 2, 2, 2, 1, 2, 2$ пута, па је њен индекс коинциденције

$$\frac{4\times3 + 2\times1 + 2\times1 + 2\times1 + 1\times0 + 2\times1 + 2\times1}{15\times14} = \frac{22}{210} \approx 0.105.$$

Код дужих текстова на истом језику, дакле и на српском, учестаности појављивања појединих слова пропорционалне су дужини текста, па се може радити са вероватноћама слова.

У табели 4 приказане су ове вероватноће израчунате у мастер раду [2] на основу текста од $50\,000$ слова. Ако је q_i вероватноћа појављивања i-тог слова азбуке, $i = 0, 1, \dots, n-1$, онда је просечни индекс коинциденције текстова на том језику једнак $\sum_{i=0}^{n-1} q_i^2$. Специјално, узимајући у обзир вероватноће из табеле 4, добија се да је индекс коинциденције текстова на српском језику приближно $\mathbf{0.064}$. С друге стране, индекс коинциденције (хаотичног, нечитљивог!) текста у коме се сва слова појављују са једнаком вероватноћом $1/n$ једнак је $n\left(\frac{1}{n}\right)^2 = 1/n$. Специјално, за такав хаотичан текст на српској ћирилици, индекс коинциденције једнак је $1/30 \approx \mathbf{0.033}$.

Чињеница да се ове две вредности, $\mathbf{0.064}$ и $\mathbf{0.033}$ толико разликују, омогућује да се израчунати индекс коинциденције текста искористи за аутоматско разликовање српског и хаотичног текста.

А	0.1212	**З**	0.0175	**Њ**	0.0074	**Ф**	0.0017
Б	0.0155	**И**	0.0930	**О**	0.1015	**Х**	0.0063
В	0.0373	**Ј**	0.0294	**П**	0.0295	**Ц**	0.0065
Г	0.0181	**К**	0.0344	**Р**	0.0479	**Ч**	0.0115
Д	0.0391	**Л**	0.0300	**С**	0.0502	**Џ**	0.0002
Ђ	0.0023	**Љ**	0.0049	**Т**	0.0445	**Ш**	0.0123
Е	0.0916	**М**	0.0362	**Ћ**	0.0058		
Ж	0.0054	**Н**	0.0551	**У**	0.0435		

Табела 4. Вероватноће слова у типичном ћиричном тексту на српском језику.

Важно је запазити да је за шифру транслације индекс коинциденције поруке једнак индексу коинциденције шифрата. Ако се на пример транслацијом слово "А" пресликава у слово "П", онда је број појављивања слова

"А" у поруци једнак броју појављивања слова "П" у шифрату, па се бројиоци индекса коинциденције поруке и шифрата састоје од истих сабирака.

Одређивање дужине кључа

Ова чињеница може се искористити за **одређивање дужине m кључа** $k_1, k_2, \dots, k_m$ Вижнерове шифре. Приметимо да се фиксирано слово k_i кључа користи за добијање истом транслацијом подниза шифрата $C_i = c_i, c_{i+m}, c_{i+2m}, \dots$ од одговарајућег низа поруке $P_i = p_i, p_{i+m}, p_{i+2m}, \dots$.

Због тога је индекс коинциденције подниза C_i исти као индекс коинциденције P_i и једнак приближно 0.064. С друге стране, ако l није умножак m, онда је подниз шифрата $c_i, c_{i+l}, c_{i+2l}, \dots$ личи на хаотичан текст, јер се од одговарајућих слова поруке добија различитим транслацијама; дакле, очекује се да овакав подниз има мањи индекс коинциденције, близак 0.033.

Према томе, дужина кључа може се одредити тако што се за вредности рецимо $l = 1, 2, 3, \dots, 15$ израчунају индекси коинциденције поднизова $c_i, c_{i+l}, c_{i+2l}, \dots, i = 1,2, \dots l$. У случају када је $l = m$, односно када се погоди дужина кључа, очекује се да израчунати индекси коинцидеције буду велики, односно блиски вредности 0.064. За остале вредности l очекује се да индекси коинциденције поднизова буду мали, блиски 0.033.

Пример 3. (Наставак примера 2) Применом описаног поступка добија се табела са вредностима индекса коинциденције различитих поднизова шифрата из примера 2 за $1 \leq i \leq l \leq 10$.

l i	1	2	3	4	5	6	7	8	9	10
1	0.042									
2	0.041	0.043								
3	0.041	0.044	0.041							
4	0.04	0.042	0.043	0.044						
5	0.07	0.061	0.073	0.065	0.061					
6	0.041	0.044	0.037	0.043	0.043	0.045				
7	0.041	0.041	0.039	0.04	0.05	0.04	0.042			
8	0.036	0.045	0.047	0.035	0.041	0.04	0.037	0.049		
9	0.038	0.044	0.041	0.043	0.044	0.035	0.04	0.045	0.044	
10	0.061	0.06	0.069	0.061	0.041	0.08	0.06	0.075	0.07	0.083

У табели су истакнуте вредности индекса коинциденције веће од 0.06. Закључује се да је дужина кључа вероватно 5, што смо и очекивали.

Чињеница да се велике вредности индекса коинциденције добијају и за претпостављену дужину кључа 10 је разумљива. Ако је дужина кључа

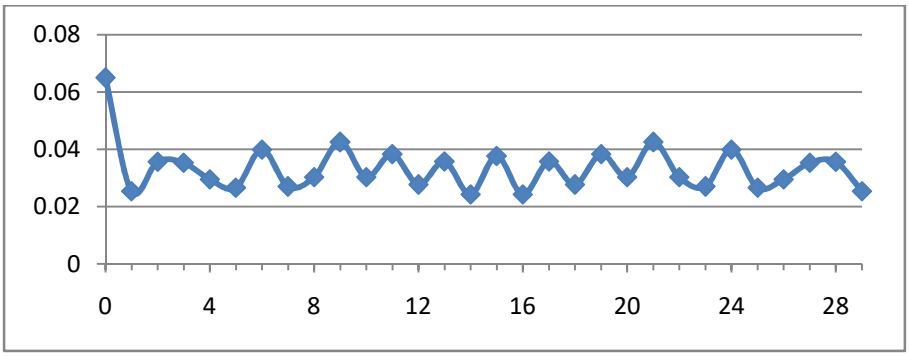
5, онда се и поднизови шифрата за размак $l = 10$ такође добијају од поруке истом транслацијом, односно шифровањем истим словом кључа.

Одређивање кључа

Пошто је одређена дужина кључа, за одређивање самог кључа може да се искористи међусобни индекс коинциденције (МИК) два текста, тј. вероватноћа да су једнака два случајно изабрана слова, једно из првог, а друго из другог текста. Ако су оба текста на српском језику, онда њихов МИК има велику вредност, блиску 0.064. Слична вредност МИК добија се за два шифрата добијена транслацијом текста на српском језику за исту вредност. С друге стране, МИК за текст на српском и текст на српском транслиран за k, $0 \leq k \leq n-1$, има вредност приближно

$$g(k) = \sum_{i=0}^{n-1} q_i \, q_{i-k \bmod n}$$

где су q_i вероватноће појединих слова. На слици 1 приказана је зависност $g(k)$ од k. Запажа се да израз има највећу вредност за $k = 0$, а да за остале k важи $g(k) \leq 0.425$.



Слика 1. Зависност међусобног индекса коинциденције текста на српском језику и његове транслације за $k = 0, 1, \ldots, 29$.

Посматрајмо сада за фиксирани индекс i подниз шифрата $C_i = c_i, c_{i+m}, c_{i+2m}, \ldots$ добијен транслацијом истим словом кључа k_i. Нека је $|C_i|$ дужина подниза C_i и нека су f_0, f_1, ... , f_{n-1} учестаности појављивања свих слова (од "А" до "Ш" ако је у питању ћирилица) у поднизу C_i. Ако са C_i^d означимо текст који се од подниза C_i добија транслацијом за $-d$ (дешифровањем C_i d-тим словом азбуке), онда је вероватноћа појављивања j-тог слова азбуке у тексту C_i^d једнака $f_{j+d \bmod n}/|C_i|$. МИК текста на српском језику и текста C_i^d и једнак је

$$\frac{1}{|C_i|}\sum_{j=0}^{n-1} q_j\, f_{j+d-k_i \bmod n} \approx g(k_i - d \bmod n) \qquad (1)$$

Овај израз има највећу вредност, блиску 0.064, ако је $d = k_i$; за остале вредности d вредности овог израза су мање, највише око 0.425. Дакле, израчунавањем ових d вредности и проналажењем највеће међу њима, можемо да одредимо свако слово кључа k_i.

Пример 4. (Наставак примера 3) Пошто је установљено да је дужина кључа једнака 5, за $i = 1, 2, 3, 4, 5$ израчунавају се вредности израза (1) за свако $d = 0, 1, \dots, 29$. Највећа од тих вредности је индекс слова кључа k_i:

i	Највећа вредност (1)	Одговарајуће d	k_i
1	0.066	18	"п"
2	0.060	6	"е"
3	0.069	21	"т"
4	0.065	0	"а"
5	0.060	19	"р"

Дакле, кључ којим је шифрована порука је реч "петар". Занимљиво је, да ако бисмо за дужину кључа изабрали 10, овим поступком добили бисмо да је кључ реч "петарпетар", што је такође тачно.

Литература

[1] P. Gannon, *Colossus*: *Bletchley Park's Greatest Secret*, Atlantic Books, 2015.

[2] А. Драгановић, *Криптоанализа Вижнерове шифре*, мастер рад, Математички факултет, 2009. http://elibrary.matf.bg.ac.rs/handle/123456789/1914

[3] М. Живковић, *Криптографија* (скрипта), Математички факултет, 2022. http://poincare.matf.bg.ac.rs/~ezivkovm//nastava/kripto.pdf

[4] D. Kahn, *The Codebreakers*, The New American Library, 1973.

[5] W. Friedman*, Military Cryptanalysis,* Part I, 4th Edition, NATIONAL SECURITY AGENCY, 1952.

[6] H. Henderson, *Encyclopedia of Computer Science and Technology*, Infobase Publishing, 2009.