KRYPTOGRAFIE ELIPTICKÝCH KŘIVEK (1)

Eliptické křivky a šifrování

Kryptografie eliptických křivek (ECC) je moderní a nadějný směr, který v řadě ukazatelů přináší lepší výsledky než současné běžně používané kryptosystémy. Seznámíme vás s podstatou i možnostmi v této oblasti.

Určitě jste se už (přinejmenším v Chipu) setkali s kryptosystémem RSA, takže asi víte, že ho můžete použít jak k šifrování, tak i k elektronickému podpisu. Že však naše vyhláška k zákonu o elektronickém podpisu umožňuje podepisovat i pomocí eliptických křivek, je fakt téměř neznámý – nejspíš i proto, že málokdo ví, co to vlastně je. V tomto dvoudílném příspěvku se pokusíme záhadu co nejpřístupnější formou objasnit.

Začínáme „od Adama"

Kdo narazí na pojem „eliptická křivka" poprvé, určitě se zeptá, co mají elipsy společného se šifrováním. Samotné elipsy sice mnoho ne, ale od nich je jen krůček k eliptickým integrálům, funkcím a křivkám. Matematikové si s těmito pojmy hráli tak dlouho, až, ostatně jako obvykle, z toho vyšlo něco praktického. Z čisté algebry vzniká po více než sto letech výzkumu „do šuplíku" v roce 1985 (pracemi V. Millera a N. Koblitze) zcela nová perspektivní oblast moderní aplikované kryptografie. Kdyby tak zakladatelé tohoto směru věděli, že se jejich myšlenky dostanou až na jakési bezkontaktní čipové karty a do čipů pro jakési technologie GSM!

V současné době pronikly eliptické kryptosystémy do řady světových standardů a staly se alternativou ke „klasickému" RSA i DSA. Mají své výhody zejména v rychlosti a menší náročnosti na hardware i software. Proč se tedy nepoužívají masově? Jednoduše proto, že „staré" kryptosystémy RSA, DSA, Diffie-Hellman, ElGamal atd. jsou používány, studovány a známy déle a mají vybudovánu infrastrukturu. Proto jsou vývojářům a technologům bližší.

Nasazení eliptických kryptosystémů vyžaduje více novátorství, obecně však umožňuje realizovat kryptografické funkce na hardwaru s nižším výkonem, o něco rychleji atd. Naproti tomu se určitě musí více investovat do začátků projektů – do výzkumu a vývoje –, a to právě kvůli novosti těchto nástrojů. Pro mnohé je tu dosud mnoho neznámého a neobvyklého. Přesto eliptické kryptosystémy už koexistují s klasickými a začínají se postupně hlásit o své místo na slunci.

Z grafu je to jasné...

Nejprve si trochu připomeneme algebru a seznámíme se s „reálnou" eliptickou křivkou (E) danou obecnou rovnicí $y^2 = x^3 + ax + b$. Musíme si ale zvyknout na to, že v algebře se dají dělat i trochu neobvyklé kousky, třeba sečíst dva body na rovinné křivce. (Nelekejte se, nakonec uvidíte, že je to vlastně velmi jednoduché.)

Na obrázku 1 vidíme konkrétní eliptickou křivku danou rovnicí $y^2 = x^3 - 4x$. Je to množina bodů (x, y) v rovině, jejichž souřadnice vyhovují uvedenému vztahu; jak vidíte, křivka má dvě oddělené části. Nyní vezmeme dva různé body P, Q, které na křivce leží, a definujeme to, co je na první pohled divné: součet bodů P + Q. Součtem bude opět bod křivky E a vznikne takto (viz obr. 1): Spojíme body $P = (x_P, y_P)$ a $Q = (x_Q, y_Q)$ přímkou; ta protne křivku v dalším bodě, který označíme -R, a výsledkem sčítání je bod R, symetrický k -R podle osy x. Body symetrické podle osy x nazýváme *opačné*.

Algebraicky je směrnice přímky, která spojuje body P, Q (zatím uvažujeme, že jsou různé a nikoli opačné), rovna $s = (y_Q - y_P)/(x_Q - x_P)$ a souřadnice bodu $R = (x_R, y_R)$ lze pak odvodit z rovnice křivky jako $x_R = s^2 - x_P - x_Q$ a $y_R = s(x_P - x_R) - y_P$. V případě P = Q přechází jejich spojnice v tečnu ke křivce E a její směrnice je rovna $s = (3x_P^2 + a)/(2y_P)$.

Když sčítáme body opačné, tj. P = -Q, měli bychom dostat „nulu", „nulový bod". Geometrická interpretace nám ovšem ihned prozradí, že spojnice takových bodů (tj. rovnoběžka s osou y) eliptickou křivku E už v žádném dalším bodě neprotne, respektive ji protne jakoby v nekonečnu. (Pro hloubavé poznamenejme, že každá jiná, byť jen nepatrně „šikmá" přímka už křivku dříve či později protne; na detailní rozbor zde však není místo.) Aby všechno fungovalo jednotně, matematici prostě definitoricky ke křivce E „bod v nekonečnu" *O* přidali, a sčítání dodefinovali i pro body opačné: P + (-P) = *O*.

Bod v nekonečnu je oficiální název onoho „nulového bodu" křivky E a musíme dodefinovat operaci sčítání i pro něj. Uděláme to přirozeně, jak se na „nulový" bod sluší: pro každý bod P na křivce definujeme P + *O* = P a také *O* + *O* = *O*, přičemž -O = O. Tím máme definováno sčítání pro všechny dvojice bodů na křivce E i bod *O*.

A nyní obecněji...

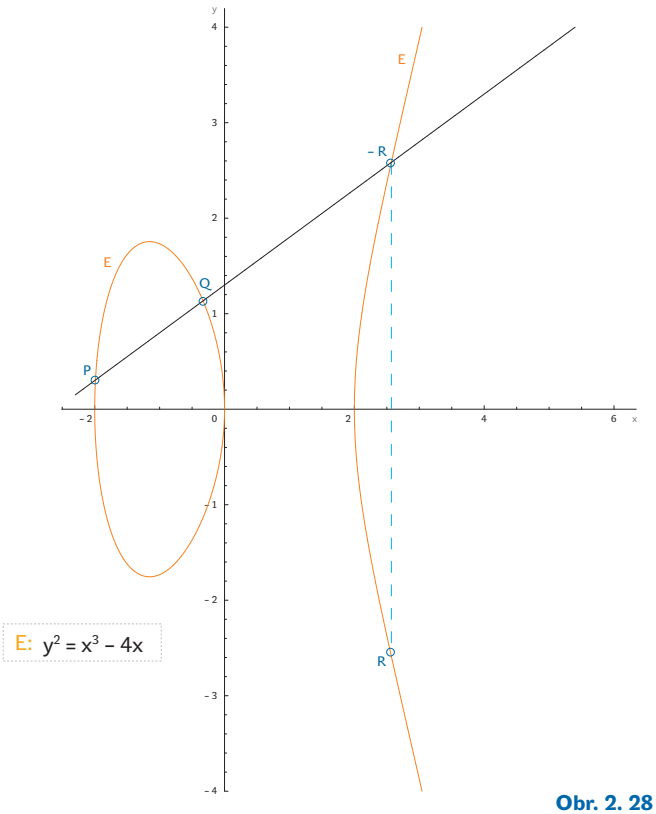
Nahraďme teď slovo křivka „vznešenějším" termínem ***grupa*** (je to prostě nějaká skupina bodů) a místo sčítání říkejme ***grupová operace*** – máme tak definovanou ***grupu E***. Protože jsme pracovali v rovině, byla souřadnicemi bodů na křivce (x, y) reálná čísla. Reálná čísla jsou takzvaným *tělesem*, což je sice další nový algebraický pojem, ale klidně na něj můžeme zapomenout. Stačí nám zde vědět jen to, že nenulovými prvky tělesa můžeme *dělit*. Podíváme-li se totiž na vzorce pro výpočet souřadnic bodu R, vidíme, že tam dělení potřebujeme (a navíc i sčítání, resp. odečítání, což jsou operace, které má těleso také k dispozici).

Proč ale vůbec potřebujeme nějaké „těleso" a nezůstaneme u reálných čísel? Pokud totiž budeme chtít něco šifrovat, budeme se muset pohybovat v oblasti diskrétních hodnot (celých čísel, bitových řetězců), a nikoli reálných čísel, protože si nemůžeme dovolit „zaokrouhlování". Proto si těleso reálných čísel nahradíme jiným tělesem (F), vhodným pro počítačové zpracování. Pak souřadnice (x, y) bodů na eliptické křivce E budou prvky tohoto tělesa F, a nikoli reálná čísla. Řečeno odborně, dostali jsme tak eliptickou křivku ***nad tělesem*** F.

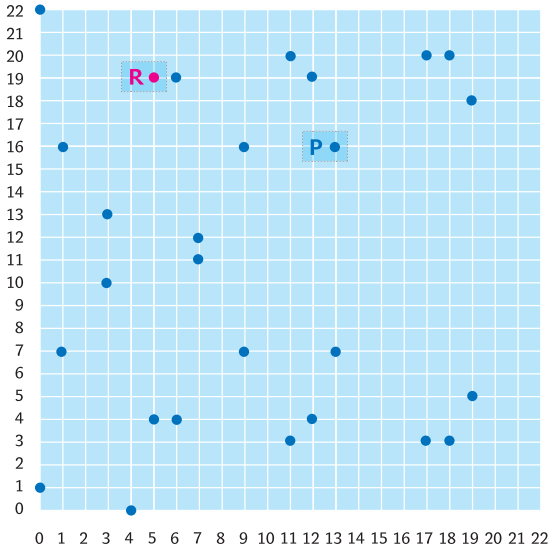
Pro počítačové zpracování je nejpřirozenější pracovat s m-ticemi bitů. To nám umožní těleso F označované jako $GF(2^m)$, které právě obsahuje všechny m-tice bitů. Dalším vhodným kandidátem F je těleso GF(p), kde p je prvočíslo. Toto těleso obsahuje čísla {0, 1, ..., p-1}, kde p je obvykle velmi velké prvočíslo, a výpočty v něm se provádějí *modulo p*. Pro další výklad si vezmeme právě těleso GF(p), protože bude o něco jednodušší než u $GF(2^m)$. Obě dvě tělesa jsou ale v praxi používána, neboť každé z nich má své přednosti.

Těleso GF(p)

Pro ty, kteří už zapomněli, co znamená „modulo p", nejprve stručné připomenutí. V GF(p) můžeme *běžně* pracovat s celými čísly, tj. násobit, sčítat, odčítat apod., jen výsledné číslo (c) se na závěr *moduluje*



Obr. 1. Sčítání bodů na rovinné eliptické křivce



Obr. 3. Body eliptické křivky $y^2 = x^3 + x + 1$ nad GF(23)

(0, 1)	(6, 4)	(12, 19)	(0, 22)
(6, 19)	(13, 7)	(1, 7)	(7, 11)
(13, 16)	(1, 16)	(7, 12)	(17, 3)
(3, 10)	(9, 7)	(17, 20)	(3, 13)
(9, 16)	(18, 3)	(4, 0)	(11, 3)
(18, 20)	(5, 4)	(11, 20)	(19, 5)
(5, 19)	(12, 4)	(19, 18)	*O*

Obr. 2. 28 bodů eliptické křivky $y^2 = x^3 + x + 1$ nad GF(23)

číslem p, což označujeme jako „c mod p“, a není to nic jiného než (celočíselný nezáporný) **zbytek po dělení** čísla c číslem p. I když GF(p) neobsahuje záporná čísla, budeme s nimi běžně pracovat, ale pokud například napíšeme -1, ve skutečnosti máme na mysli „-1 mod p“, čili p-1. Všechna čísla se zkrátka chápou „modulo p“.

Jak jsme viděli v předchozích vzorcích pro sčítání bodů na křivce, v našem tělese ještě budeme potřebovat operaci dělení. Definujeme ji podobně jako u reálných čísel prostě jako násobení inverzním číslem, například x/y je $x*(y^{-1})$ a přirozeně y^{-1} je ten prvek tělesa GF(p), který vynásoben y dává jedničku: $y*(y^{-1}) = 1$. Abychom zdůraznili, že se jedná o násobení v tělese GF(p), píšeme uvedenou rovnici ve formě $y*(y^{-1}) \equiv 1 \pmod p$.

Například v GF(23), tj. pro p = 23, máme $5^{-1} = 14$, protože $14*5 = 70 = 23*3 + 1$, a tudíž $14*5 \equiv 1 \pmod{23}$. Je to trochu divné, ale funguje to. Tak třeba výpočet 4 + (3-7)/5 uděláme jako $4 + (-4)*5^{-1} = 4 + (-4)*14 = 4 - 56 = -52 = (-3)*23 + 17$, čili vyjde 17.

Teď už tedy umíme počítat v tělese GF(p) a můžeme si uvést rovnici eliptické křivky nad GF(p). Uvidíme, že to bude zcela analogické jako u reálných čísel.

Eliptická křivka E nad tělesem GF(p)

Eliptická křivka E nad tělesem GF(p) je definována jako bod v nekonečnu *O* společně s množinou bodů P = (x, y), kde x a y jsou z tělesa GF(p) a splňují rovnici $y^2 = x^3 + ax + b$ v GF(p), tj. $\mathbf{y^2 \equiv x^3 + ax + b \pmod p}$.

Koeficienty a, b v rovnici jsou také prvky tělesa GF(p) a musí splňovat podmínku (pro jednoduchost ji uvádíme bez odvození) $4a^3 + 27b^2 \pmod p \neq 0$, která zaručuje, že takto definovaná množina bodů tvoří grupu (jinak koeficienty *a* a *b* můžeme volit libovolně – budou to později veřejné parametry příslušného kryptosystému). V této grupě definujeme opačný bod k *O* jako $-O = O$ a pro ostatní nenulové $P = (x_P, y_P) \in E$ definujeme $-P = (x_P, -y_P \bmod p)$, dále pro všechny body $P \in E$ definujeme $P + -P = O$ a $P + O = P$. Bod *O* nazýváme také nulový bod, vzhledem k jeho roli při sčítání v grupě E. Sčítání stejných nenulových bodů P + P definujeme jako $R = P + P = (x_R, y_R)$, kde směrnice s je rovna $s = (3*x_P^2 + a)/(2*y_P)$ a souřadnice $x_R = s^2 - x_P - x_P$, $y_R = s(x_P - x_R) - y_P$. Sčítání různých nenulových a vzájemně neinverzních bodů $P = (x_P, y_P)$ a $Q = (x_Q, y_Q)$ definujeme jako $R = P + Q = (x_R, y_R)$, kde směrnice s je rovna $s = (y_Q - y_P)/(x_Q - x_P)$ a souřadnice $x_R = s^2 - x_P - x_Q$, $y_R = s(x_P - x_R) - y_P$.

Všimněte si, že tato definice je totožná s definicí sčítání bodů na eliptické křivce nad reálným tělesem. Takže ve skutečnosti nic nového, jen se všechny operace provádějí modulo p. Ukážeme si to ještě na příkladu.

Příklad

Zvolme p = 23, a = 1, b = 1. Protože $4a^3 + 27b^2 \pmod{23} = 4 + 27 \pmod{23} \neq 0$, máme tak eliptickou křivku $E: y^2 = x^3 + x + 1$ nad GF(23). Její body jsou vypsány na obrázku 2 a graficky znázorněny na obrázku 3. Můžeme si například ověřit, že bod (0, 1) patří této křivce: platí totiž $1^2 \equiv 0^3 + 0 + 1 \pmod{23}$.

2P

Nyní podle definice sečteme body P + P, kde P = **(13, 16)**. Máme $R = P + P = (x_R, y_R)$, kde $s = (3*13^2 + 1)/(2*16) = 508/32 = 508*18 = 9144 = 13$. (Snad vás příliš nepopudil zkrácený zápis, který jsme právě začali používat; např. podivná rovnost 9144 = 13 platí samozřejmě jen při výpočtech modulo 23.) Dále dostáváme $x_R = 13^2 - 13 - 13 = 143 = 5$, $y_R = 13*(13 - 5) - 16 = 88 = 19$, tj. R = **(5, 19)**; viz obr 3. Výsledek krátce označíme 2P, což je symbolický zápis pro P + P.

3P

Vypočtěme ještě R = P + P + P = (P + P) + P = 2P + P (výsledný bod označme opět symbolicky 3P). Sčítáme tedy body (5, 19) a (13, 16): $s = (16 - 19)/(13 - 5) = -3/8 = -3*3 = 14$, $x_R = 14^2 - 5 - 13 = 178 = 17$, $y_R = 14*(5 - 17) - 19 = 20$, takže 3P = **(17, 20)**.

- 4P, 5P, 6P, 7P

Podobně bychom vypočítali 4P = (17, 3), 5P = (5, 4), 6P = (13, 7) a 7P = *O*. Bod 7P můžeme obdržet jako 7P = 6P + P = (13, 7) + (13, 16) = (13, 7) + (-(13, 7)) nebo třeba jako 4P + 3P = (17, 3) + (17, 20), v obou případech se jedná o sčítání opačných bodů, takže výsledkem je bod *O*. Při výpočtu 8P se pochopitelně dostaneme zpět k bodu P, neboť 8P = 7P + P = *O* +P = P.

Svinujeme graf...

Eliptická křivka nad tělesem GF(23) na obr. 3 nám naši známou reálnou křivku už geometricky příliš nepřipomíná. Určitě je však zřejmá symetrie kolem pomyslné osy y = p*1/2 = 11,5. To je zákonité, protože na křivce leží vždy jak bod (x, y), tak bod k němu opačný, tj. (x, -y), což je (x, p-y), tedy právě body symetrické podle uvedené osy y = p*1/2. (Pokud se na rozsypané body v mříži podíváte velmi pozorně, naleznete tam náznak původní reálné křivky s tím, že osa x je nyní posunuta do polohy y = p*1/2, tj. nahoru o 11,5.) Převedení křivky nad celočíselné těleso zkrátka boří jednoduché vztahy, což je vlastně celý záměr kryptografie eliptických křivek, neboť to, co mohlo být na reálné křivce řešitelné jednoduše, bude na diskrétní křivce složité.

Ale uvažujme dále. Podívejme se na rovinnou eliptickou křivku z obrázku 1, zkusme si v její rovině představit celočíselnou síť a uvědomit si, že v této rovině počítáme modulo p, tj. původní křivku nad tělesem reálných čísel jakoby převedeme na křivku nad tělesem GF(p).

Kryptografie eliptických křivek je moderní a nadějný směr, přinášející výsledky v řadě ukazatelů lepší než stávající kryptosystémy.

Protože počítáme modulo p, splývají v GF(p) všechny reálné body typu ..., (x, y - 2*p), (x, y - p), (x, y), (x, y + p), (x, y + 2*p), ..., protože jejich y-ové souřadnice budou totožné. Je to tedy, jako bychom rovinu srolovali podélně podle osy x tak, aby obvod ruličky byl roven p, a y-ové souřadnice jsme měřili na povrchu této ruličky od nuly do p (pak uvedené body splývají v jeden). Na ruličce máme nyní body (x, y), kde y je jen z intervalu 0 až p, zatímco x probíhá celou reálnou osu.

Protože i souřadnice x je v GF(p), body typu ..., (x - 2*p, y), (x - p, y), (x, y), (x + p, y), (x + 2*p, y), ... jsou totožné. Můžeme tedy naši ruličku ještě stočit do kolečka tak, aby jeho obvod byl roven p – pak uvedené body také splynou. Výsledkem je prstencový útvar zvaný torus nebo anuloid; nejlépe si jej představíme jako nafouknutou duši do pneumatiky.

Zkusme teď domyslet, jak bude na toru vypadat původní reálná křivka E. Při prvním svinování roviny do ruličky se nám křivka (resp. její neuzavřená část) začne spirálovitě kroutit po ruličce směrem doprava do nekonečna a při druhém stočení ruličky do kolečka se tento běh křivky promítne na torus tak, že křivka bude po něm nekonečně rotovat. Torus bude proto dost „počmáraný“, nás ovšem zajímají jen ty body křivky, které na jeho povrchu protnou celočíselnou síť, která má za souřadnice celá čísla od nuly do p-1. Průnik reálné křivky s touto celočíselnou sítí na toru pak ve skutečnosti vidíme na obrázku 3. Když přimhouříte oči, uvidíte, že obrázky 1 a 3 (byť se jedná o různé křivky) mají přece jen hodně společného.

Diskrétní logaritmus v eliptické křivce

Konečně se dostáváme k šifrování a podepisování na eliptické křivce. Jde o využití tzv. *problému diskrétního logaritmu*. Uvidíme, že má velmi jednoduchou podstatu. Vezměme si třeba bod P = (13, 16) ∈ E, definovaný v předchozím příkladu, a vypočítejme postupně 2P, 3P, 4P, 5P, 6P atd., čímž dostáváme obecně různé body (xP) na křivce. Protože křivka má konečný počet bodů (označíme ho #E), po určitém počtu kroků (m) se nám musí tato posloupnost zacyklit. V bodě zacyklení (mP) tak platí mP = nP, kde nP je nějaký dřívější bod. Odtud ale dostáváme mP - nP = (m-n)P = *O*, čili existuje nějaké r (= m - n < m) takové, že rP = *O*, takže je jasné, že v posloupnosti P, 2P, 3P, 4P, ... se vždy nakonec dostaneme do bodu *O* a poté cyklus začíná znovu od bodu P, neboť (r+1)P = rP + P = *O* + P = P.

Nejmenší takové r, pro něž je rP = *O*, nazýváme ***řád bodu P***; v našem příkladu jsme viděli, že bod P = (13, 16) měl řád r = 7. Lze dokázat, že řád bodu dělí řád křivky, přičemž ***řádem křivky*** nazýváme počet bodů (#E) na křivce (v našem příkladu je #E = 28 = 2^2*7). Různé body na křivce E mohou mít kupodivu různý řád. V kryptografické praxi vybíráme takový bod, jehož řád je roven největšímu prvočíslu v rozkladu čísla #E, nebo jeho násobku, který nazýváme *kofaktor*; například r = 2*7 má kofaktor 2.

U bodu řádu r máme tedy zaručeno, že v posloupnosti P, 2P, 3P,..., dojde k zacyklení až po r-tém kroku. Pokud je r velké, například řádově 2^{256}, je to opravdu velmi dlouhá posloupnost. Právě při šifrování a elektronickém podepisování takovéto velké posloupnosti využíváme, a to v souvislosti s tzv. problémem diskrétního logaritmu. Zvolíme tajné číslo *k* (je to náš *privátní klíč*) a vypočteme bod Q = *k*P. Body P a Q můžeme nyní zveřejnit – budou součástí našeho *veřejného klíče*. (Zveřejníme samozřejmě i popis křivky E.)

Problémem diskrétního logaritmu (DLP, *discrete logarithm problem*) je právě úloha, jak z bodů P a Q určit ono tajné číslo *k* tak, že Q = *k*P. Pro malý řád bodu P je tato úloha triviální, pro velká r je to však úloha, kterou matematici už nedovedou efektivně (tj. v polynomiálním čase) řešit. Proto body P a Q můžeme zveřejnit, protože z nich nelze tajné číslo *k* v dohledné době určit. Dosud nejúčinnější metodou na řešení této úlohy je tzv. Pollardova ró metoda (v Chipu jsme o ní psali v jiné souvislosti, viz [2] a [3]), jejíž složitost je řádově $(\pi*r/2)^{1/2}$ kroků. Pokud je r = 2^{256}, dostáváme cca 2^{128} kroků, což je zhruba na úrovni luštitelnosti symetrické blokové šifry se 128bitovým klíčem, a pro nás z výpočetního hlediska neřešitelné. Proto říkáme, že příslušná šifra je výpočetně bezpečná.

Ukázalo se sice, že úlohu lze paralelizovat, takže je to podobné jako paralelizace hledání klíče u symetrických šifer, jak bylo vidět například u DES-Crackeru (viz literatura). Pokud použijeme N procesorů, dostáváme složitost $(\pi*r/2)^{1/2}$ / N, ale pro velká r je to stále výpočetně neřešitelná úloha.

Úlohu DLP si teď ukážeme na naší eliptické křivce. Zveřejníme body P = (13, 16) a Q = (17, 20). Jaké je *k*? Odpověď je *k* = 3, jak je vidět z příkladu. Pro malý řád bodu P tuto úlohu zvládneme – ale co když je cca 2^{160}...?

Základy máme, příště šifrujeme...

V tomto článku jsme se seznámili s rovnicí eliptické křivky nad konečným tělesem a s tzv. problémem diskrétního logaritmu. V příštím pokračování si ukážeme, jak se problém diskrétního logaritmu dá využít k definici perspektivních kryptosystémů s veřejným klíčem.

■ ■ ■ Vlastimil Klíma, *autor@chip.cz*

Literatura

[1] Klíma, V.: DES Cracker, Chip 11/98, str. 74 – 75, a 12/98, str. 52 – 54, všechny články jsou dostupné též elektronicky na [4]

[2] Klíma, V.: Dvě čísla za 200 000 dolarů, Chip 9/01, str. 176 – 180

[3] Rosa, T.: Podpis k narozeninám, Chip 8/01, str. 131 – 133

[4] Archiv článků: *http://www.decros.cz/bezpecnost/_kryptografie.html*