běžně se volí r = 0,01 nebo r = 0,001. Pokud vypočtená hodnota S padne do uvedeného intervalu spolehlivosti, přijmeme hypotézu, že daná posloupnost má maximální entropii. Pokud S padne mimo něj, hypotézu zamítneme. V tom případě ji ale zamítneme správně, neboť tak činíme s pravděpodobností 1 - r, tj. téměř s jistotou. Pro obě obvykle volené hodnoty r jsou příslušné IS uvedeny v tabulce 1, stejně jako obecný vzorec.

Pokud tedy například pro L = 8 obdržíme S = 7,995, můžeme přijmout hypotézu, že se jedná o náhodný zdroj s maximální entropií. Obdržíme-li S = 4,002, hypotézu odmítneme, ale pokud bylo zdrojových dat velké množství, můžeme učinit závěr, že každých 8 bitů produkované posloupnosti obsahuje v průměru cca 4 bity neurčitosti.

Pro ilustraci použitelnosti a schopností M-C testu jsme udělali několik experimentů, jejichž výsledky uvádí tabulka 2. Samozřejmě je vždy lepší data testovat s větším rozlišením testu (při L = 16 je test mnohem přesnější než při L = 8), ale připomeňme, že pro L = 16 test vyžaduje minimálně 65 MB zdrojových dat, zatímco pro L = 8 stačí jen 256 KB.

V prvních čtyřech experimentech jsme použili krátké texty, v pátém jeden dlouhý text. Přesto na nich test nefunguje ani zdaleka tak dobře jako běžně dostupný komprimační program WinZip. Proč? Text totiž není pro M-C test vhodný. Text má hluboké závislosti, které zakladatel teorie informace C. E. Shannon ve svých raných pracích odhadoval (v angličtině) i při zjednodušeném modelu minimálně na pět znaků (jinými slovy, výskyt písmene čitelného textu závisí až na pěti předchozích písmenech). Museli bychom proto měřit entropii pro L = 5*8 = 40 bitů, což by vyžadovalo text o délce přes $1000*2^{40}$ znaků, tj. 1000 terabajtů. I tak bychom ale nezaregistrovali takové zákonitosti a opakování textu, jaké zachytí i běžný komprimační program. Jeho „okno“ totiž bývá nikoli pět, ale až 8000 znaků.

Z experimentů 6 až 8 je vidět, že L = 16 poskytuje přesnější měření než L = 8, a M-C test se blíží výsledkům WinZipu. Je to víceméně náhodný výsledek, protože uvedené formáty vlastní zdroj dat samy zásadně upravují a přetvářejí ve velmi velkých blocích, takže se jedná o zdroje dat značně umělé. Zjišťovat u nich míru entropie by vyžadovalo studovat jejich systém kódování vstupních dat do výsledného formátu a odlišit skutečné vstupy od přídavných rámců nebo formátovacích sekvencí. Měřit u nich entropii M-C testem je proto nesmyslné.

V dalších dvou experimentech jsme pro zajímavost změřili entropii pohybu myši ze vzorku jejích poloh, pořízených asi za 10 sekund (experiment 9), pokud jsme jí záměrně pohybovali, a za jednu hodinu běžné práce u PC (experiment 10), tj. včetně doby, kdy se používá více klávesnice a občas myš, tj. kdy se většinu času nepohybuje. WinZip je zde lepší, protože poloha myši se zapisuje prostřednictvím 32 bitů, které M-C test s L = 8 a 16 nemůže tak dobře vyhodnotit.

Následující experimenty už jdou M-C testu „k duhu“. Abychom mohli vyhodnotit účinnost testu na velkém množství dat, zvolili jsme zdroj dat s entropií 1,000000, tj. zcela náhodný zdroj dat, poté binární zdroj s entropií 0,937500 na bit (15 bitů ze 16 je náhodných, poslední bit je dopočítán jako paritní) a potom binární zdroj s entropií 0,875000 na bit (14 bitů z 16 je náhodných, 15. bit je paritní bit za předchozí liché bity a 16. bit je paritní za předchozí sudé bity). Na těchto souborech WinZip zcela odmítl komprimovat, neboť je považoval za náhodné a nedosáhl žádné komprimace. To by bylo v pořádku u skutečně náhodných souborů v experimentech 11 a 14, ale ne už u ostatních, kde měl komprimovat na cca 93 % a 87 %. WinZip tam ale neodhalil žádnou zákonitost, zatímco M-C test zapracoval fantasticky přesně. Třeba v experimentech 14 až 16 jím zjištěné entropie **1,000000**, **0,937511** a **0,875008** jsou až neuvěřitelně blízko skutečným entropiím měřených zdrojů. Tyto výsledky také ukazují, že na přirozených náhodných zdrojích je M-C test velmi přesný, a čím menší paměť zdroj má (nejlépe když následné hodnoty jsou zcela nezávislé), tím je přesnější.

Dále se zde ukazuje další možné použití M-C testu. Pokud data mají nějakou závislost v okně délky N bitů, M-C test to nezjistí pro parametr L < N, ale při L = N a větší ano (srv. testy 11 až 16 pro L = 8 a L = 16). Jestliže máme podezření, že předložená data takovou zákonitost skrývají, lze ji odhalit provedením všech tes-

Tabulka 1. Intervaly spolehlivosti a další parametry.

Interval spolehlivosti (t1,t2) pro entropii S pro různá L

		ρ = 0,01			ρ = 0,001		
L	σ	y	t1	t2	y	t1	t2
6	0,00437		5,98873	6,01127		5,98561	6,01439
7	0,00315		6,99188	7,00812		6,98963	7,01037
8	0,00226		7,99419	8,00581		7,99258	8,00742
9	0,00161		8,99586	9,00414		8,99472	9,00528
10	0,00114	2,57583	9,99706	10,00294	3,29053	9,99625	10,00375
11	0,00081		10,99792	11,00208		10,99734	11,00266
12	0,00057		11,99853	12,00147		11,99812	12,00188
13	0,00040		12,99896	13,00104		12,99867	13,00133
14	0,00029		13,99926	14,00074		13,99906	14,00094
15	0,00020		14,99948	15,00052		14,99933	15,00067
16	0,00014		15,99963	16,00037		15,99953	16,00047

Poznámka 1:
K je pevně voleno jako $1000*(2^L)$
a interval spolehlivosti se vypočítá podle vztahu
$(t1, t2) = (L - y*\sigma, L + y*\sigma)$,
kde σ je dáno v tabulce a y se vypočítá z ρ prostřednictvím vztahu N(-y) = ρ,
kde N je distribuční funkce normálního rozdělení, tj.

$$N(x) = (1/2\pi)^{1/2} * \int_{-\infty}^{x} e^{-\xi^*\xi/2}\, d\xi,$$

L je délka bloku a σ je dáno tabulkou.

Poznámka 2:
Intervaly spolehlivosti pro L = 8 a L = 16, vztažené na 1 bit (údaje v tabulce, dělené L):
L = 8: <0,9992737, 1,0007262> a <0,9990725, 1,0009275>
L = 16: <0,9999768, 1,0000231> a <0,9999706, 1,0000293>

Trocha filozofie

tů pro L = 4, 5, 6, ..., N, N+1, ... Pokud obdržené hodnoty S budou vykazovat v bodě L = N zásadní zlom, máme už jistotu, že N-bitové vzorky nějakou neznámou zákonitost obsahují, a můžeme se pokusit ji odhalit. To je další výsledek, který je hodnotný sám o sobě.

Maurerův-Coronův test je účinný na testy fyzikálních a svým charakterem přírodních (originálních) zdrojů informace. Není vhodný na měření entropie umělých generátorů, například kongruentních nebo kryptografických posloupností. Vysvětlíme si to na příkladu. Mějme třeba zdroj, který má entropii 0,5 na jeden bit výstupu. Dále uvažujme, že máme tajnou substituční tabulku (8 bitů na 8 bitů), kterou aplikujeme na každý bajt originální posloupnosti. Pokud použijeme M-C test s délkou bloku L = 8, pak entropie původní i modifikované posloupnosti budou naprosto totožné!

Sebetajnější substituce výsledek neovlivní, neboť test nezajímají konkrétní hodnoty znaků, ale jejich vztahy, ale ty se substitucí nemění. Kdybychom použili 128bitovou substituci (např. blokovou šifru), museli bychom u M-C testu volit také 128bitové bloky (tj. L = 128), abychom její vliv eliminovali. M-C test by v tomto případě vyžadoval zpracování $1000*2^{128}$ bloků, což je ale výpočetně nezvládnutelné. Jinými slovy, pokud výstupní posloupnost nemá dostatečnou entropii, nemá cenu ji uměle doupravovat a pak měřit entropii upravené posloupnosti M-C testem. Je ale možné volit obrácený postup. Ze zdroje, který má M-C testem zjištěnou určitou entropii, nejprve generujeme posloupnost, až dosáhneme požadované entropie, a teprve poté tuto posloupnost můžeme upravovat, abychom získanou entropii využili.

Tabulka 2. Výsledky experimentů.

Aplikace Maurerova-Coronova testu entropie na různé typy souborů a porovnání s komprimačním programem WinZip							
č.	Typ souboru	Koncovka	Velikost	Hodnota entropie na 1 bit zdroje získaná M-C testem	Testová charakteristika M-C testu a poznámka	Hodnota entropie na 1 bit získaná komprimací WinZip	Poznámka
1	český text s háčky a čárkami	txt	15 KB	0,742322	L = 8 *	0,454256	text článku [4]
2	anglický text	txt	15 KB	0,704996	L = 8 *	0,402977	anglický překlad [4]
3	český text v MS Wordu	doc	75 KB	0,498773	L = 8 *	0,234321	text článku [4]
4	anglický text v MS Wordu	doc	64 KB	0,425613	L = 8 *	0,208338	anglický překlad článku [4]
5	anglický text v ASCII	txt	30 MB	0,544607	L = 8	0,246832	normy RFC jako jeden text
				0,444148	L = 16 *		
6	bitmapový soubor	bmp	1,4 MB	0,185204	L = 8	0,011122	obrázek 2
				0,090877	L = 16 *		
7	český text v Adobe Acrobatu	pdf	0,3 MB	0,977951	L = 8	0,917281	[4]s obrázky
8	Chip CD 12/98	pdf	27 MB	0,965729	L = 8,	0,890385	obsah celého čísla
				0,922664	L = 16 *		
9	záznam pohybu myši	dat	5,8 KB	0,751005	L = 8	0,709137	pohyb myši za 10 s
10	záznam pohybu myši	dat	0,3 MB	0,344991	L = 8	0,039239	pohyb myši cca za 1 hod. práce u PC
				0,111001	L = 16 *		
11	binární zdroj s entropií 1,0/bit	rng	66 MB	1,000728	L = 8	větší než 1,0	WinZip nekomprimuje, data natahuje
				0,999994	L = 16		
12	binární zdroj s entropií 0,9375 / bit	rng	66 MB	0,999449	L = 8		
				0,937507	L = 16		
13	binární zdroj s entropií 0,875/bit	rng	66 MB	0,996907	L = 8		
				0,874999	L = 16		
14	binární zdroj s entropií 1,0/bit	rng	132 MB	1,000702	L = 8		
				1,000000	L = 16		
15	binární zdroj s entropií 0,9375/bit	rng	132 MB	0,999434	L = 8		
				0,937511	L = 16		
16	binární zdroj s entropií 0,875/bit	rng	132 MB	0,996907	L = 8		
				0,875008	L = 16		

* = nekorektně krátká délka dat

Závěr

Maurerův-Coronův test je univerzální test náhodnosti, který je schopen detekovat širokou škálu statistických defektů. Na jejich odhalení není pak nutné používat další speciální statistické testy. Kromě toho test přímo poskytuje číselný odhad entropie daného zdroje. Může být využit k měření entropie přirozených zdrojů, kde je nutná záruka kvality nebo znalost míry jejich náhodnosti, nehodí se ale k testování umělých zdrojů ani tam, kde jsou tyto zdroje uměle upravovány.

VLASTIMIL KLÍMA (*V.KLIMA@DECROS.CZ*)

Literatura

[1] Maurer, U., „An Universal Statistical Test for Random Bit Generators“, Proceedings of CRYPTO´90, Lecture Notes in Computer Science, pp. 409 – 420, Springer-Verlag, 1990.

[2] Coron, J. S., Naccache, D., „An Accurate Evaluation of Maurer's Universal Test“, Proceedings of SAC'98, Lecture Notes in Computer Science, Springer-Verlag, 1998.

[3] Coron, J. S., „On the Security of Random Sources“, Public Key Cryptography, Lecture Notes in Computer Science, vol. 1560, pp. 29 – 42, Springer-Verlag, 1999

[4] Klíma, V., „Až nás podepíše počítač“, Chip 5/99, str. 36 – 39.