Testy a zdroje neurčitosti v počítači

Dá se náhoda měřit?

V oblasti počítačové bezpečnosti se velmi často setkáváme s náhodnými čísly a šifrovacími klíči. Na kvalitě “náhodnosti” jejich generování přitom záleží úplně stejně jako na kvalitě používaných šifer. V tomto článku vás seznámíme s nedávným objevem, který umožňuje měřit kvalitu náhodnosti daného zdroje. Je to poměrně přesná metoda, jejíž význam však sahá daleko za hranice počítačové bezpečnosti. Možná vás už nějaký program požádal, abyste chvíli náhodně ťukali do klávesnice nebo pohybovali myší. To jsou okamžiky, kdy na náhodnosti záleží natolik, že program odmítá za kvalitu svého zdroje převzít odpovědnost a obrací se přímo na uživatele. Znát míru náhodnosti používaného zdroje je nutné zejména u bezpečnostních aplikací. Kritické je to pak při generování šifrovacích klíčů. Jestliže generátor náhodných bitů nemá dostatečnou kvalitu, může se stát, že vygenerovaných 128 bitů šifrovacího klíče má pouze 40bitovou informační hodnotu (neurčitost, *entropii*). Generátor tak může snadno degradovat silnou šifru na slabou a důsledky mohou být značné. Tyto případy se už staly – a bohužel určitě nikoli naposledy.

Bezpečnost a náhoda

Přestože kvalitní zdroj entropie je při práci na počítači potřeba dost často, s požadavkem vložení náhodného čísla se v praxi setkáváme málokdy. Příslušné programy totiž nechtějí obtěžovat uživatele a generují náhodnost samy – jak umějí nejlépe. Ve většině případů k tomu využívají pouze “náhodnost” odvozenou od systémového času, což je ale z hlediska bezpečnosti silně nedostatečné. Náhodné šifrovací klíče musí například generovat internetový prohlížeč, pokud se se serverem spojuje zabezpečeným spojením prostřednictvím protokolu SSL. Jak možná víte, starší verze prohlížeče Netscape Navigator používala slabý generátor náhodných čísel, a šifrovací klíče tak měly entropii 47 namísto 128 bitů. Tím se degradovala kvalita šifrování a byla z toho ostuda.Od té doby se na kvalitu náhodných generátorů dbá více.

Komprimace a náhodnost

Entropie vlastně určuje skutečné množství obsažené informace a měří se v bitech. Jednoduchým a známým měřítkem náhodnosti mohou proto být např. komprimační metody. Pokud nějaký soubor dat zkomprimujeme dejme tomu na 40 % původní délky, můžeme říci, že 60 % obsahu bylo nadbytečných a skutečný informační obsah byl 40 %. V jednom bajtu bylo tedy obsaženo jen 40 %, tj. 8*0,4 = 3,2 bitu skutečné informační hodnoty (entropie), neboli průměrná entropie na jeden bit byla 0,40. A co komprimovaný soubor – bude náhodný? Téměř ano, i když na jeho začátku mohou být prvopočáteční kusy původního textu a v jeho těle některé markantní řetězce. V mnoha případech ale komprimace skutečně velmi přiblíží soubor dat jeho informační hodnotě. Jestliže ale dáme zkomprimovat soubor náhodných dat, komprimační metody zkolabují. A to i v případech, že zdrojová data nejsou zcela náhodná, ale mají entropii například 0,90. Komprimace by měla daný soubor zkrátit na 90 %, ale nestane se tak, protože příslušná metoda prostě nezjistí, o jakou neurčitost vlastně jde. Neumí ji zjistit, změřit ani odstranit. V případech náhodných nebo téměř náhodných souborů tedy běžné komprimační metody jako měřítko neurčitosti použít nelze.

Objev v měření entropie

Průlom v měření entropie znamenal objev Ueliho Maurera z roku 1990, který jej prezentoval na kryptologické konferenci CRYPTO´90 [1]. Nalezl velmi jednoduchou funkci, jíž dokázal měřit a pomocí statistického testu testovat entropii generátoru. Do té doby byla známa řada důmyslných testů, které zkoumaly partikulární parametry posloupnosti, jako například statistické vlastnosti (autokorelační test, test sérií, frekvenční test apod.) nebo složitostní charakteristiky, ale výsledky se nedaly kvantitativně převést na hodnotu entropie. Jinými slovy – věděli jsme, že posloupnost dejme tomu 200 pozic myši (měřených v časových mikrointervalech při jejím pohybu, viz obrázek 2) není náhodná a jaké má nedostatky (korelace sousedních pozic, nerovnoměrný výskyt jednotlivých bajtů), ale nevěděli jsme, jak dlouho máme myší pohybovat, aby posloupnost jejích pozic už reprezentovala například 128bitovou entropii, tj. ekvivalent 128 náhodných bitů. A Maurerův test právě toto dokázal vypočítat.

Použitelnost Maurerova-Coronova testu

V roce 1999 zpřesnil odhady konstant Maurerova testu J. S. Coron [2] a poté navrhl i geniální změnu testovací funkce [3]. Nový test tak oproti dřívějšímu měří entropii přímo a přesněji. Pro vás, kteří byste jej chtěli přímo použít, jej dále popíšeme. Test se týká stacionárních zdrojů s konečnou pamětí. Přesné definice a důkazy tvrzení můžete nalézt v uvedené literatuře. “Stacionární” znamená, že se v čase nemění charakteristiky zdroje (například na pohyb myši nemá vliv to, zda je měřen v úterý, nebo ve čtvrtek), a konečná paměť (M) znamená, že n-tý výstup zdroje závisí maximálně jen na konečném počtu (M) předchozích výstupů – například poloha myši v daném okamžiku závisí maximálně na tom, kde byla před sekundou, ale už ne na tom, kde byla před deseti sekundami.

Výpočet entropie

Pojďme tedy k výpočtu entropie S podle Maurerova-Coronova (dále jen M-C) testu. Nejprve si zvolíme tři parametry – konstanty L, Q a K. Testovanou posloupnost N bitů si dále rozdělíme na Q + K nepřekrývajících se L-tic bitů b_1, ..., b_{K+Q}, kde b_i je i-tý blok o L bitech a N = (Q + K)*L.

Parametr L by měl být volen v rozmezí {6, ..., 16}, Q by mělo být co největší, minimálně ovšem $10*2^L$ a K alespoň $1000*2^L$. Jestliže např. zvolíme L = 8, zpracováváme posloupnost po bajtech.

Test má dvě fáze – inicializační a výpočetní. V **inicializační fázi** nejprve naplníme tabulku T[0] ... T[2^L-1] indexy prvních Q bloků tak, že pro i = 1, ..., Q postupně definujeme T[b_i] = i. Jinými slovy: prvních Q bloků použijeme na to, abychom naplnili tabulku T. Hodnota T[blok] je místo, kde se naposledy objevil L-bitový blok s hodnotou “blok”. Q by mělo být tak velké, aby se v inicializační fázi korektně naplnila tabulka T, tj. aby se v prvních Q blocích posloupnosti alespoň jednou objevil každý L-bitový blok.

Hodnotu S určíme ve **výpočetní fázi** podle vzorce na obrázku 3. Hodnota, kterou obdržíme, je rovna entropii L-bitového bloku. Chceme-li zjistit entropii zdroje na jeden emitovaný bit, postačí obdrženou hodnotu S vydělit počtem bitů bloku L – zdroj poskytuje neurčitost H = S/L na jeden bit. Pokud se nad vzorcem zamyslíme, zjistíme, že je to vlastně průměrná hodnota jakési funkce g, aplikované na vzdálenost mezi totožnými L-bitovými bloky v dané posloupnosti. Přitom průměr se počítá přes všechny bloky v posloupnosti.

Genialita funkce g je v tom, že uvedenou vzdálenost bloků “přeměňuje” na entropii a navíc výpočet hodnoty S je velmi jednoduchý. Coron dokázal, že **S** z teoretického hlediska **vyjadřuje hodnotu entropie přesně** – navíc **známe její statistické rozdělení**. Umíme tedy entropii zdroje nejen vypočítat, ale určit i tzv. intervaly spolehlivosti, v nichž se naměřené hodnoty S mohou pohybovat, má-li mít zdroj maximální entropii.

Hodnocení výsledků

Když použijeme M-C statistický test na zkoumanou posloupnost, obdržíme jednu jedinou hodnotu – tzv. *statistiku S*. Tato statistika je ve střední hodnotě rovna přímo entropii L-bitového bloku zkoumaného zdroje, ale zároveň je to náhodná veličina, která má pravděpodobnostní chování. A tak, i když má zdroj dokonalou entropii (například S = 8 na bajt), hodnoty S naměřené na konkrétních posloupnostech se mohou pohybovat v určitých intervalech kolem této dokonalé entropie. Tyto intervaly spolehlivosti (IS) umíme vypočítat, neboť S můžeme aproximovat normálním rozdělením, jehož parametry naposledy zpřesnil právě J. S. Coron [3].

K výpočtu IS si nejprve stanovíme pravděpodobnost **r**, že M-C testem vyřadíme nějakou posloupnost jako špatnou (nemající maximální entropii), přestože byla emitována skutečně náhodným zdrojem; 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 testů 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.

Trocha filozofie

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.

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.