

Ukážeme vylepšenou verziu algoritmu (Simon 1988)

Nechť $G = (V, E)$ je acyklický graf. Rozklad $X_1, \dots, X_k$ množiny V (tj. $X_i \neq \emptyset$ pro $1 \leq i \leq k$, $X_i \cap X_j = \emptyset$ pro $i \neq j$, $X_1 \cup \dots \cup X_k = V$) se nazývá rozklad na řetězce, jestliže každý X_i je cestou grafu G .

Pokud G je topologicky uspořádaný, máme

$$X_i = \{v_1, \dots, v_{s_i}\}, \text{ kde } (v_j, v_{j+1}) \in E \text{ pro } 1 \leq j \leq s_i.$$

Číslo k se nazývá šířka dekompozice.

Označme

$\text{id}(v)$ číslo řetězce obsahujícího v , tj. $\text{id}(v) = i \Leftrightarrow v \in X_i$

a pro množinu $A \subseteq V$

$\text{niv}_j(A) = \min(A \cap X_j)$ tj. nejmenší bod v řetězci X_j , který je kážděm prvkem A

$$\text{niv}(A) = \{\text{niv}_j(A), j = 1, \dots, k\}.$$

Speciálně pro $A = \text{out}^*(v)$ budeme psát

$$\text{niv}_j(v) = \text{niv}_j(\text{out}^*(v)), \text{ niv}(v) = \text{niv}(\text{out}^*(v)).$$

Platí:

$$\text{out}^*(v) = \bigcup_{1 \leq j \leq k} (\text{out}^*(v) \cap X_j) = \bigcup_{1 \leq j \leq k} \{w \in X_j; w \geq \text{niv}_j(v)\}$$

limba upřesnění dostaneme konkrétní ukazatel, pokud bychom uměli spočítat $\text{niv}_j(v)$

platí: $out^*(v) = \{v\} \cup \underbrace{out^*(w_1) \cup \dots \cup out^*(w_s)}_A$

tedy

$$niv_j(v) = niv_j(out^*(v)) = niv_j(\{v\} \cup A) =$$

$$= \min((\{v\} \cup A) \cap Z_j) = \min((\{v\} \cap Z_j) \cup (A \cap Z_j)) =$$

$$= \min(\underbrace{(\{v\} \cap Z_j)}_{niv_j(v)}, \underbrace{\min(A \cap Z_j)}_{niv_j(A)}) = \begin{cases} v, & \text{tedy } v \in Z_j \\ \min(niv_j(w_1), \dots, niv_j(w_s)) & \text{jinak} \end{cases}$$

platí

$$niv_j(A) = \min((out^*(w_1) \cup \dots \cup out^*(w_s)) \cap Z_j) =$$

$$= \min((out^*(w_1) \cap Z_j) \cup \dots \cup (out^*(w_s) \cap Z_j)) =$$

$$= \min(\underbrace{\min(out^*(w_1) \cap Z_j)}_{niv_j(w_1)}, \dots, \underbrace{\min(out^*(w_s) \cap Z_j)}_{niv_j(w_s)})$$

to lze se pak psát: $niv(v) = niv(out^*(v))$.

Vylepšení algoritmu spočívá v tom, že se provede

$$out^*(v) \leftarrow out^*(v) \cup out^*(w) \quad \dots \text{čas } O(n)$$

následně operaci

$$niv(out^*(v)) \leftarrow niv(out^*(v) \cup out^*(w)) \quad \dots \text{čas } O(k)$$

Obvykle iterací tj. podstatně méně než v předchozím.

Jestliže $w \in out^*(v)$ se pak dává na náhladu vlastnosti

$$w \notin out^*(w_1) \cup \dots \cup out^*(w_s) \Leftrightarrow w < \min(niv_{id(w)}(w_1), \dots$$

$$\dots niv_{id(w)}(w_s))$$

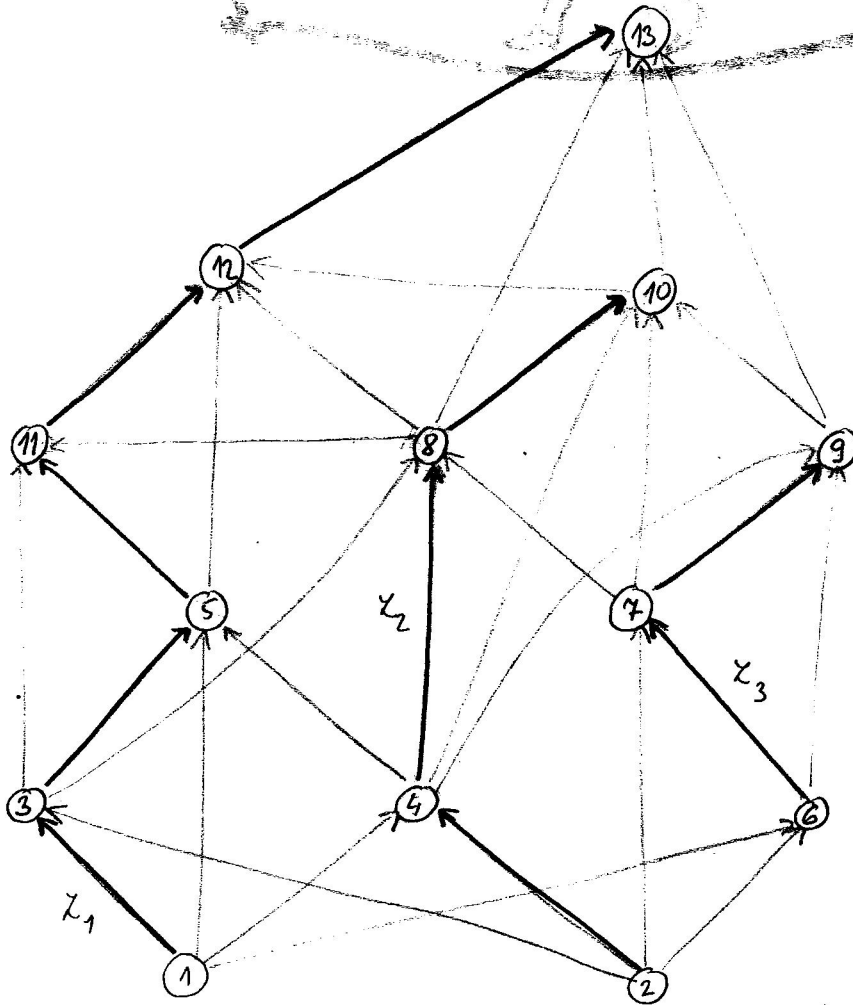
(důkaz je jednoduchý)

Postup: 1) Spočítá se relaxovaná dekompozice (hladové)

2) Pro každé $v \in V$ se spočítá $\text{niv}(v) = (\text{niv}_1(v), \dots, \text{niv}_k(v))$

3) Hledají se E^*

Příklad:



hledám $\text{out}^*(4)$: $\text{niv}_1(4) = 5$... seznam 5, 11, 12, 13

$\text{niv}_2(4) = 4$... seznam 4, 8, 10

$\text{niv}_3(4) = 9$... seznam 9

(6a)

Složitost 1) je zřejmě $O(n + |E|)$

Složitost 2) :

Řetězec, ve kterém vrchol vleží, se bere od tohoto vrcholu.

V jiném řetězci se najde vrchol takový, do kterého z v vede hrana, ale nikoli cesta délejší než 2.

Neholí dělá se průnik řetězce s tranzitivním reduktem vrcholu v .

To se dělá pro každý vrchol a každý řetězec, složitost je tedy $O(|E| + k \cdot |E^{\text{red}}|)$.

(Tranzitivní redukta se opět konstruuje tak, že se vrcholy procházejí v opačném pořadí vzhledem k topologickému uspořádání, opalované přidávání těchto vrcholů tedy nehrozí).

Složitost 3) je zřejmě $O(|E^*|)$.

Cellková složitost: $O(|E^*| + k \cdot |E^{\text{red}}|)$

Platí: $E(k) \leq \frac{\log(p \cdot n)}{p} + 1 = O\left(\frac{\log(p \cdot n)}{p}\right)$

$$E(k \cdot |E^{\text{red}}|) = O(n^2) \text{ pro } \frac{\log^2 n}{n} \leq p < 1$$
$$= O(n^2 \log \log n) \text{ jinak}$$

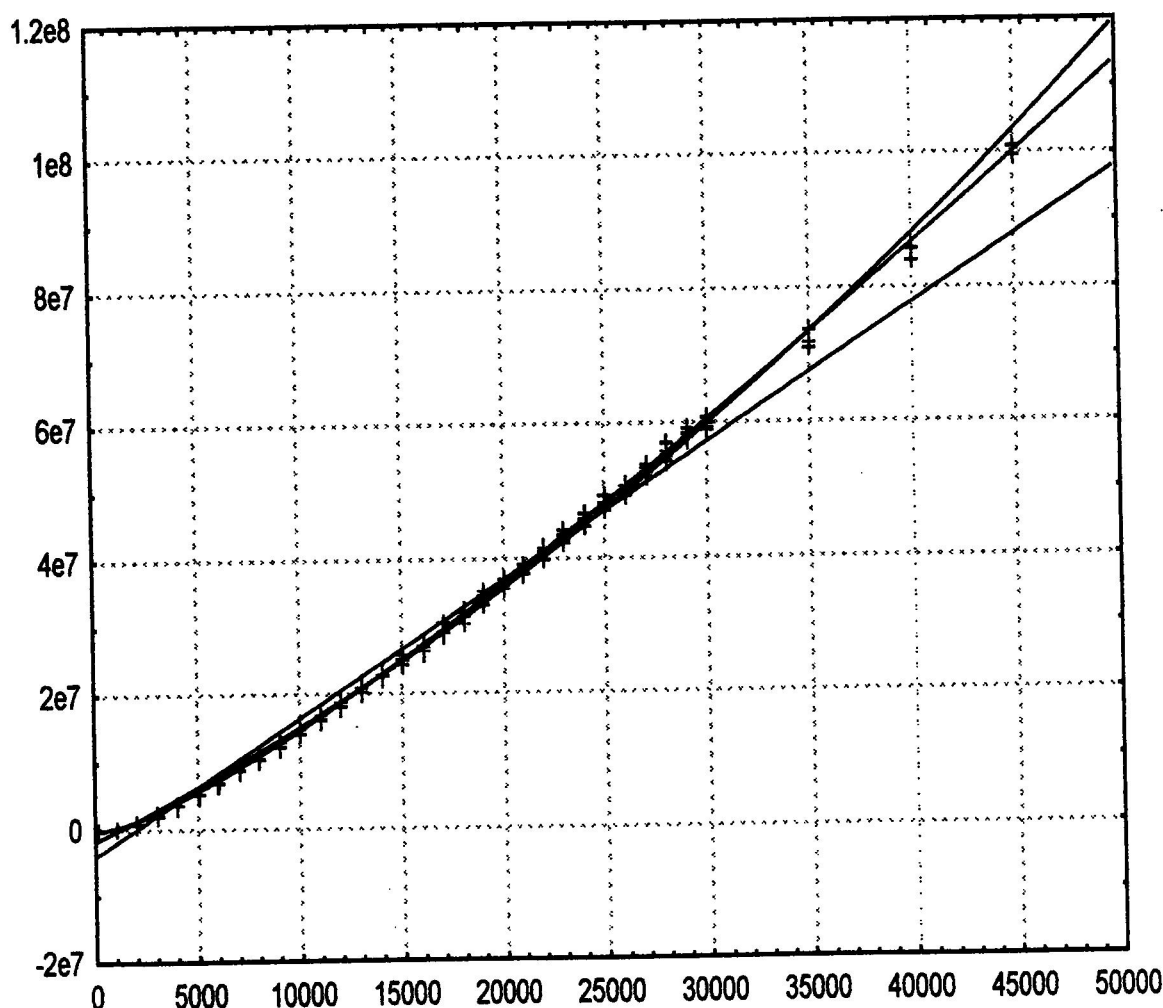
Při výpočtu pozor - veličiny k a $|E^{\text{red}}|$ nejsou nezávislé!

Příklad. Algoritmus na konstrukci tranzitivního uzdueru acyklického grafu (Simon)
(M. Černý, diplomová práce, 2003)

Následující obrázek ukazuje srovnání tří regresních modelů

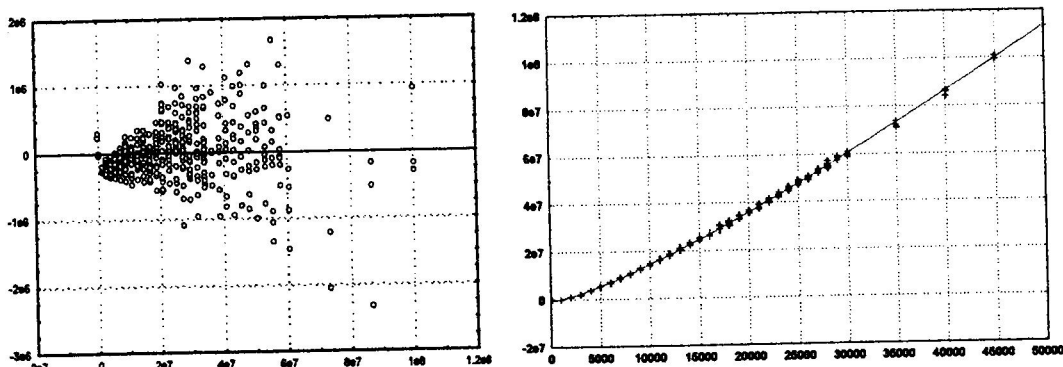
$$\begin{aligned}y &= b_1 + b_2x + b_3x^2 \quad (\text{regresní parabola}), \\y &= b_1 + b_2x + b_3x \ln x \quad (\text{model "x ln x"}), \\y &= b_1 + b_2x \quad (\text{regresní přímka})\end{aligned}$$

prokládající redundanci pro $p=0.01$. Z obrázku nelze první dva modely příliš rozlišit (kvadratický roste v obrázku nepatrně rychleji, viz graf 9.24). Tabulka 9.3. ukazuje, že v obou případech je koeficient determinace tak blízky 100%, že oba modely se jeví jako vhodné pro naše data. Lineární model se jeví jako nevhodný (nejspodnější přímka v grafu 9.24); ovšem to, že by redundance mohla růst pomaleji než kvadraticky, vyloučit nemůžeme.

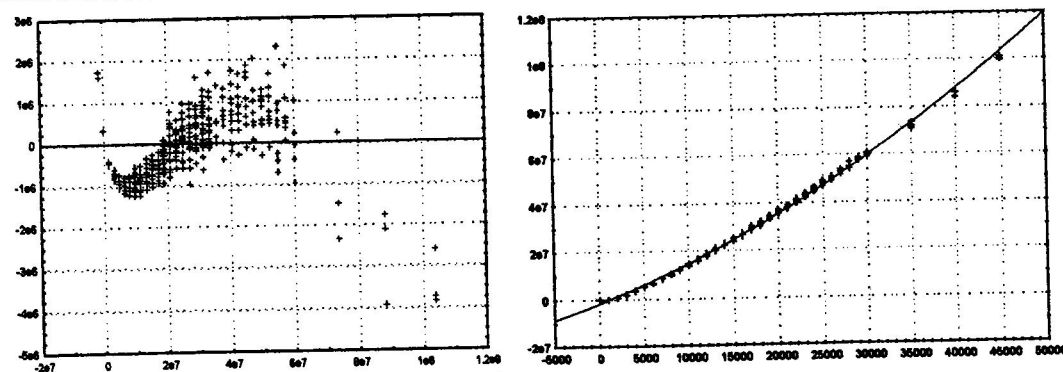


Obr. 9.24. Model kvadratický (horní, modrá křivka), $x \ln x$ (střední, zelená křivka) a lineární (spodní červená přímka) pro redundanci Simonova algoritmu $p = 0.01$

Ve prospěch modelu s členem $x \ln x$ ovšem hovoří grafy residuí. Na následujícím obrázku ukazuje první dvojice grafů residua a model $x \ln x$, zatímco druhý graf zobrazuje residua při proložení parabolou; je vidět, že parabola malé hodnoty podhodnocuje a velké hodnoty nadhodnocuje, zatímco residua modelu $x \ln x$ mají sice nekonzistentní rozptyl (což lze očekávat), ale jejich střední hodnota je blízko nuly a nejsou korelovaná, což potvrzuje i Durbin-Watson¹² (pro model $x \ln x$ je 1.79, zatímco pro parabolu 0.31) a pořadová korelace¹³ (pro model $x \ln x$ je 0.11, pro parabolu 0.84).



Obr. 9.25a. Residua a model $x \ln x$



Obr. 9.25b. Residua a model $y = b_1 + b_2x + b_3x^2$

V případě, kdybychom chtěli srovnat ještě model

$$y = b_1 + b_2x + b_3x \ln x + b_4x \ln \ln x,$$

bude již matice Z (viz kap. 8.2) natolik špatně podmíněná, že výsledky mohou být zkreslené numerickými chybami; i přesto uveďme, že vychází p-value t-testu koeficientu b_4 v řádu 0.2. Nemůžeme dost dobře z našich naměřených dat mezi modelem " $x \ln x$ " a " $x \ln \ln x$ " rozlišit.

Následující tabulka ukazuje celkový přehled regresních modelů pro redundanci Simonova algoritmu. Sloupec "p-val b_i " obsahuje p-value t-testů na regresní parametry b_1 , b_2 a b_3 , je-li

¹² viz kap. 8.4.

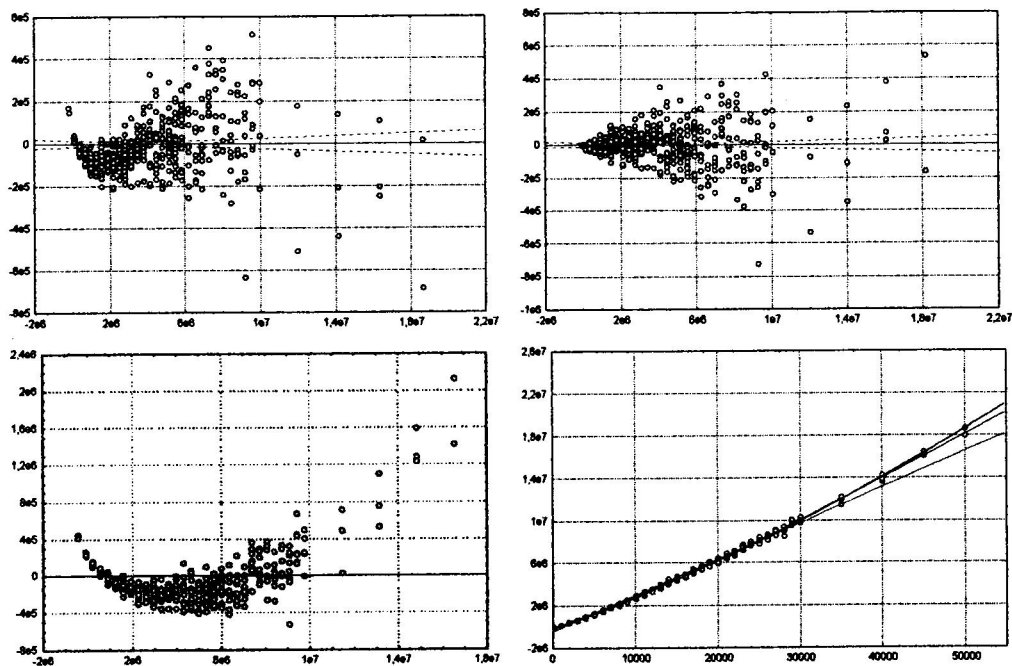
¹³ pořadová korelace je neparametrický test na korelaci v residuích, viz [4]. Hodnoty blízké 1 ukazují na silnou korelaci, hodnoty blízké nule na nekorelovanost

b_3 přítomen, "D. W." je hodnota Durbin-Watsonovy statistiky, "corr" je pořadová korelace v residuích, " R^2 " je koeficient determinace.

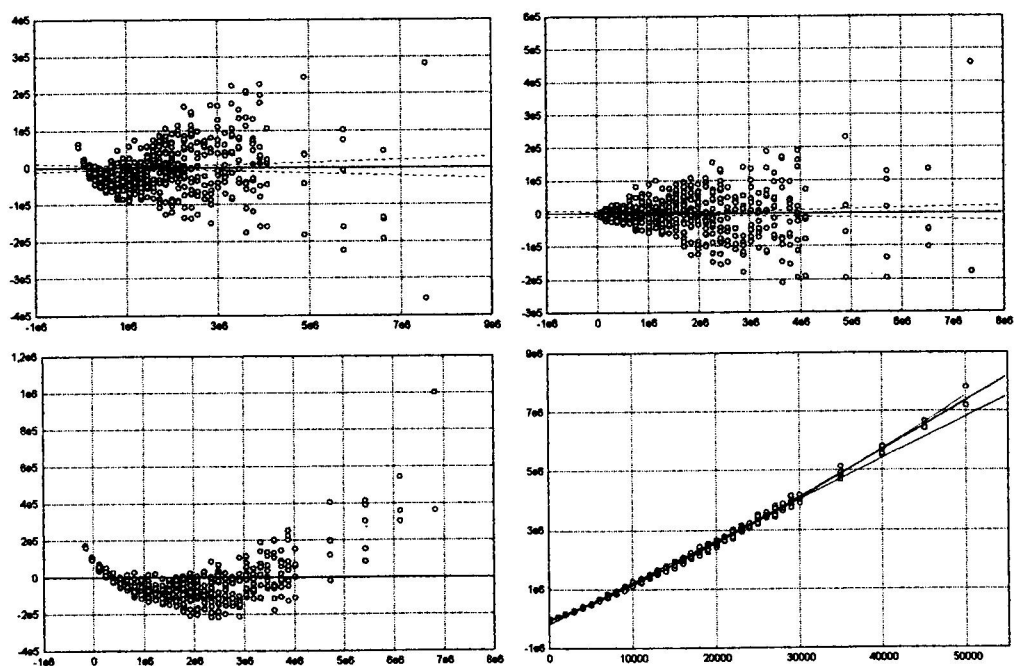
		b_1	b_2	b_3	p-val b_i	D.W.	corr	R^2
$p=0.01$	parabola	-1771191	1491.83	0.018814	0, 0, 0	0.31	0.84	99.87%
	$x \ln x$	-222457.59	-3182.46	504.7382	0, 0, 0	1.76	0.11	99.97%
	přímka	-4068100	2045.54		0, 0	0.07	0.98	99.28%
$p=0.05$	parabola	-174 449.403	279.506	0.0019443	0, 0, 0	1.22	0.38	99.91%
	$x \ln x$	-1 359.271	-230.829	54.87015	0, 0, 0	1.85	0.072	99.94%
	přímka	-442 822.305	993.9976		0, 0	0.29	0.86	99.62%
$p=0.1$	parabola	-66 038.826	116.630	0.0007104	0, 0, 0	1.61	0.179	99.88%
	$x \ln x$	1857.658	-75.44398	20.58461965	0, 0, 0	1.96	-0.03	99.9%
	přímka	-175432.489	139.958 661		0, 0	0.5	0.78	99.65%
$p=0.15$	parabola	-25505.203	63.076	0.0004740462	0, 0, 0	1.75	0.1	99.86%
	$x \ln x$	-2166.58	-26.8073	9.9796962	0.49, 0, 0	2.12	0.07	99.88%
	přímka	-64851	75.3871433		0, 0	0.83	0.57	99.71%
$p=0.25$	parabola	-13510.585	29.11927793	0.000139797	0, 0, 0	1.83	0.07	99.79%
	$x \ln x$	-1478.339	-6.1136943	3.97747859	0.42, 0.001, 0	2.06	-0.03	99.81%
	přímka	-30777,9381	33,2341103		0, 0	1.16	0.41	99.66%
$p=0.45$	parabola	-3595.23	8.646089	0.0000439721	0, 0, 0	1.66	0.15	99.55%
	$x \ln x$	-425.72	-1.398681	1.09221955	0.59, 0.1, 0	1.76	0.1	99.57%
	přímka	-8215.02	9.8526532		0, 0	1.76	0.1	99.45%
$p=0.65$	parabola	-1098.98	2.8078	0.0000149727	0, 0, 0	2.05	0.02	99.40%
	$x \ln x$	-78.83	-0.5270	0.36362218175	0.79, 0.11, 0	2.12	0.06	99.42%
	přímka	-2672.04	3.218710		0, 0	1.74	0.12	99.29%
$p=0.99$	parabola	-2.909	0.012202	0.00000018515	0.4, 0, 0	2.09	-0.04	95.34%
	$x \ln x$	4.888	-0.002109	0.003716638	0.3, 0, 0	2.09	-0.04	95.33%
	přímka	-20.37	0.01704		0, 0	1.93	0.03	94.91%

Tab. 9.3. Srovnání regresních modelů pro redundanci Simonova algoritmu

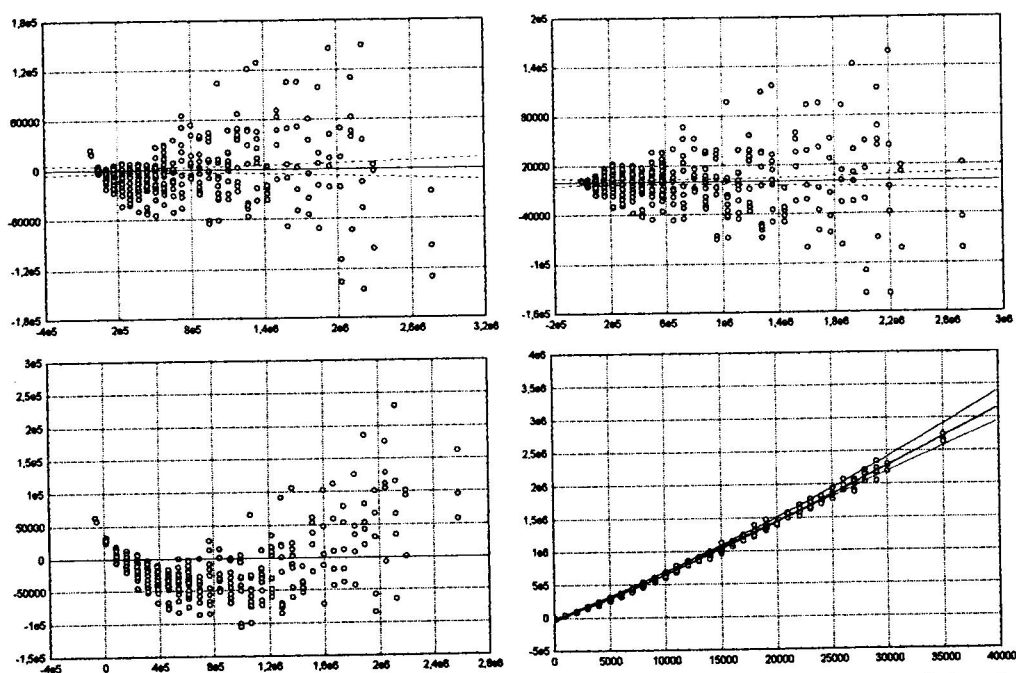
Další obrázky ilustrují údaje v tabulce. Pro každé p (s výjimkou 0.01, neboť tento případ byl diskutován v předchozím textu) uvádíme graf residuí pro všechny tři modely (parabola, $x \ln x$ a přímka) spolu obrázkem srovnání těchto modelů. Následuje vždy po řadě parabola, vedle ní model $x \ln x$, pod nimi přímka a srovnání všech tří modelů. V těch případech, kdy se jeví model jako smysluplný, uvádíme v obrázcích residuí i 95%-ní intervaly spolehlivosti pro střední hodnotu. Ve všech obrázcích srovnání modelů platí, že modrá křivka, jež ve v pravé části grafu nejvýše, je parabola, zelená střední křivka je $x \ln x$ a nejnižší položena je červená přímka.



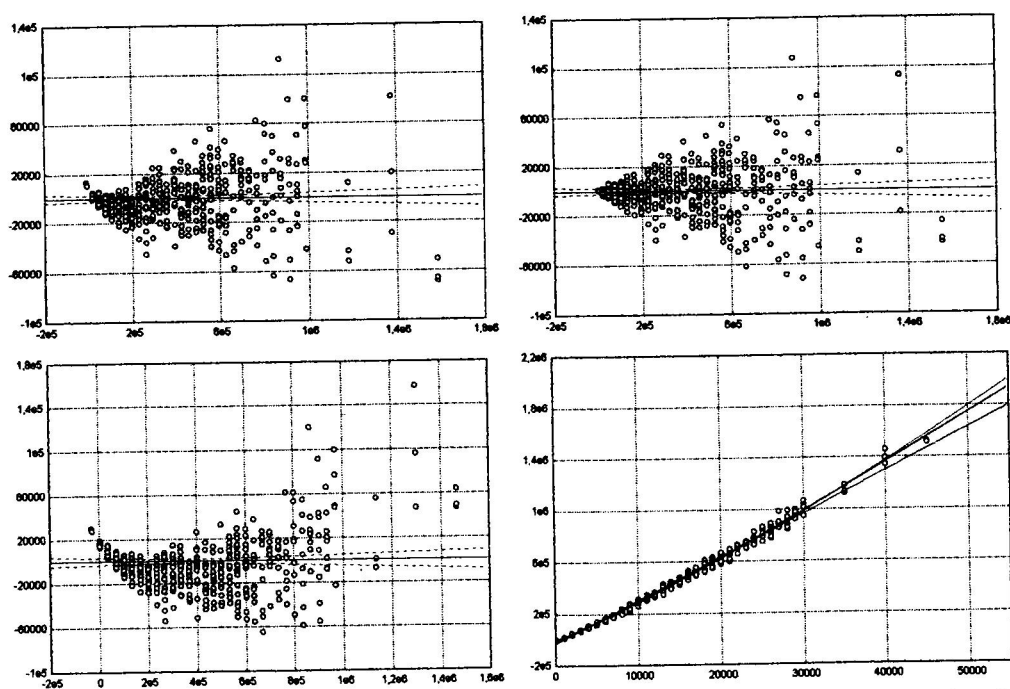
Obr. 9.26. Simonův algo, $p = 0.05$, residua pro parabolu (a), $x \ln x$ (b), přímku (c), srovnání modelů (d)



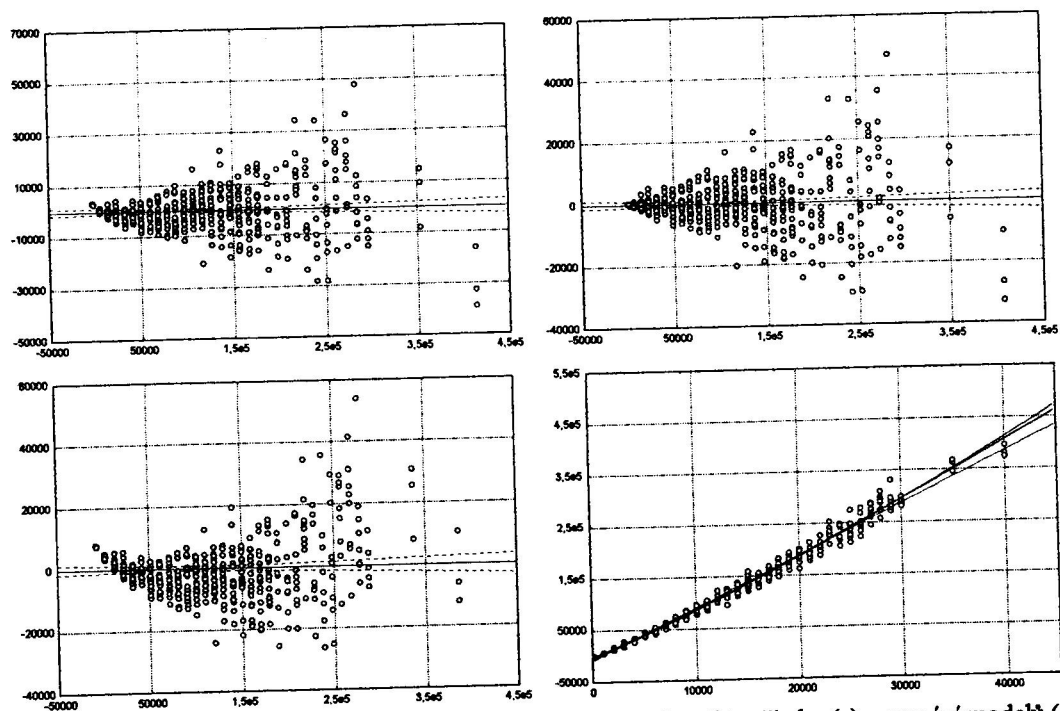
Obr. 9.27. Simonův algo, $p = 0.1$, residua pro parabolu (a), $x \ln x$ (b), přímku (c), srovnání modelů (d)



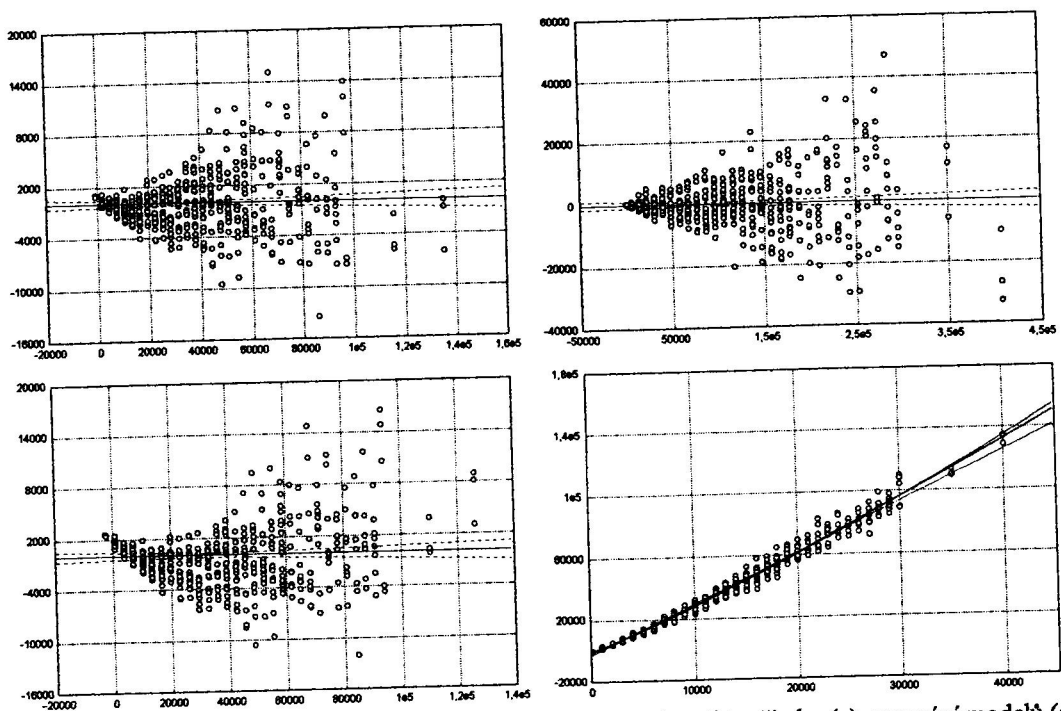
Obr. 9.28. Simonův algo, $p = 0.15$, residua pro parabolu (a), $x \ln x$ (b), přímku (c), srovnání modelů (d)



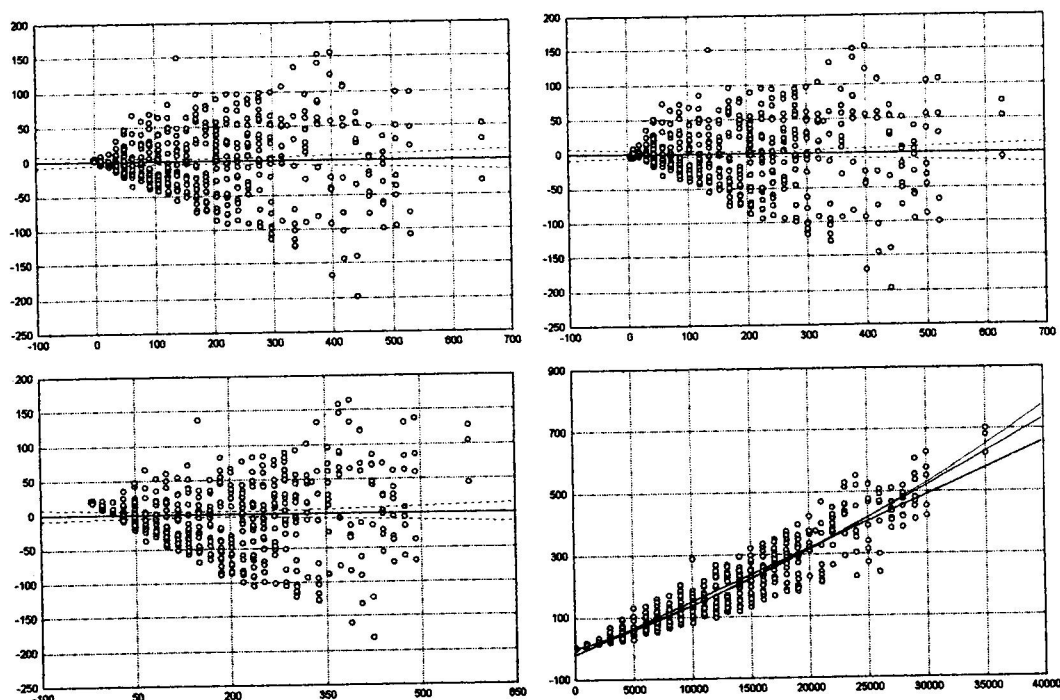
Obr. 9.29. Simonův algo, $p = 0.25$, residua pro parabolu (a), $x \ln x$ (b), přímku (c), srovnání modelů (d)



Obr. 9.30. Simonův algo, $p = 0.45$, residua pro parabolu (a), $x \ln x$ (b), přímku (c), srovnání modelů (d)



Obr. 9.31. Simonův algo, $p = 0.65$, residua pro parabolu (a), $x \ln x$ (b), přímku (c), srovnání modelů (d)



Obr. 9.32. Simonův algo, $p = 0.99$, residua pro parabolu (a), $x \ln x$ (b), přímku (c), srovnání modelů (d)

Z předchozí tabulky a obrázků se ukazuje, že ve všech případech se model $x \ln x$ jeví jako vhodný, pro menší hodnoty p je dokonce výrazně vhodnější než trend kvadratický; lineární trend se ve všech případech jeví jako nejméně vhodný. Usuzujeme z toho, že – alespoň podle provedených měření – redundance není lineární, avšak zdá se být významně subkvadratická (o tom viz i kap. 9.10).

Ve svém článku [35] Simon uvádí výsledek citovaný v kapitole 5, že pro model $G_{n,p}$ má jeho algoritmus složitost v průměrném případě $O(n^2)$, je-li $p \geq (\ln n)/n$ a $O(n^2 \ln \ln n)$ v ostatních případech. Simonův algoritmus totiž běží v čase $O(e^* + ke^R)$; je zřejmé, že člen e^* zlepšit nelze, neboť vypočtený transitivní uzávěr musí být alespoň zapsán na výstup, avšak z našich simulací vychází, že člen ke^R , který Simon odhaduje jako n^2 , resp. $n^2 \ln \ln n$, by bylo možné zlepšit dokonce až na $n \ln n$; naše simulace tedy napovídají, že pro pevné p pracuje Simonův algoritmus nad grafem $G_{n,p}$ v průměrném čase $O(e^* + n \ln n) = O(e^*)$.