Kvantitativní analýza internetového provozu (6)

Ladislav Lhotka
⟨*lhotka@cesnet.cz*⟩

Osnova přednášky

- Struktura IP adres ve vzorcích
- Fraktální dimenze
- Multifraktály

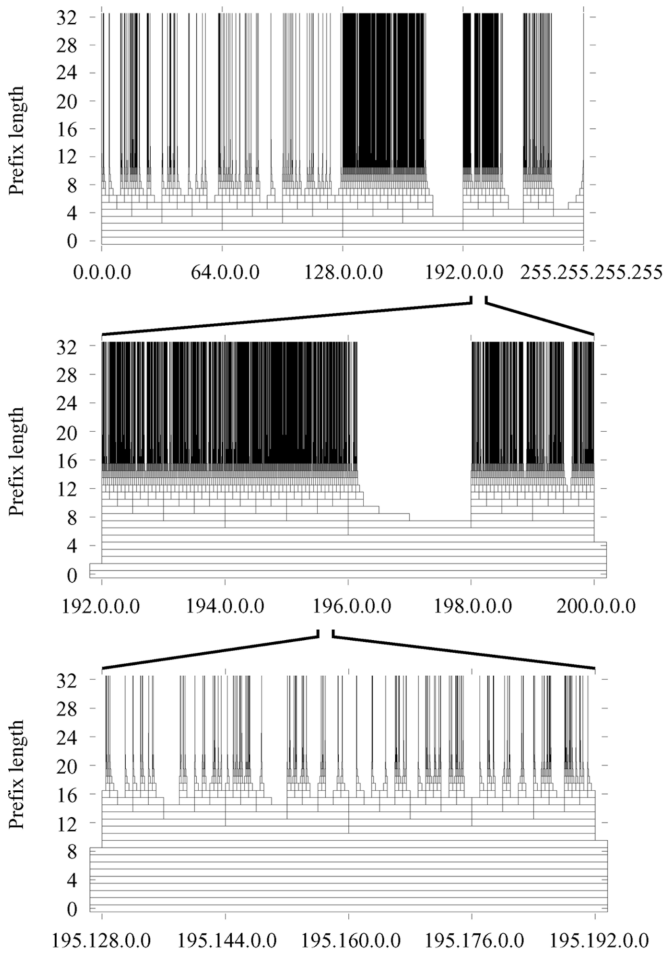
Struktura adres ve vzorcích

Kohler et al. Observed structure of addresses in IP traffic. *IEEE Transaction on Networking* 14(6), 2006, p. 1207–1218.

Adresový agregát pro prefix délky p: množina adres shodujících se v prvních p bitech. Zápis: 147.251.0.0/16, nebo jen 147.251/16.

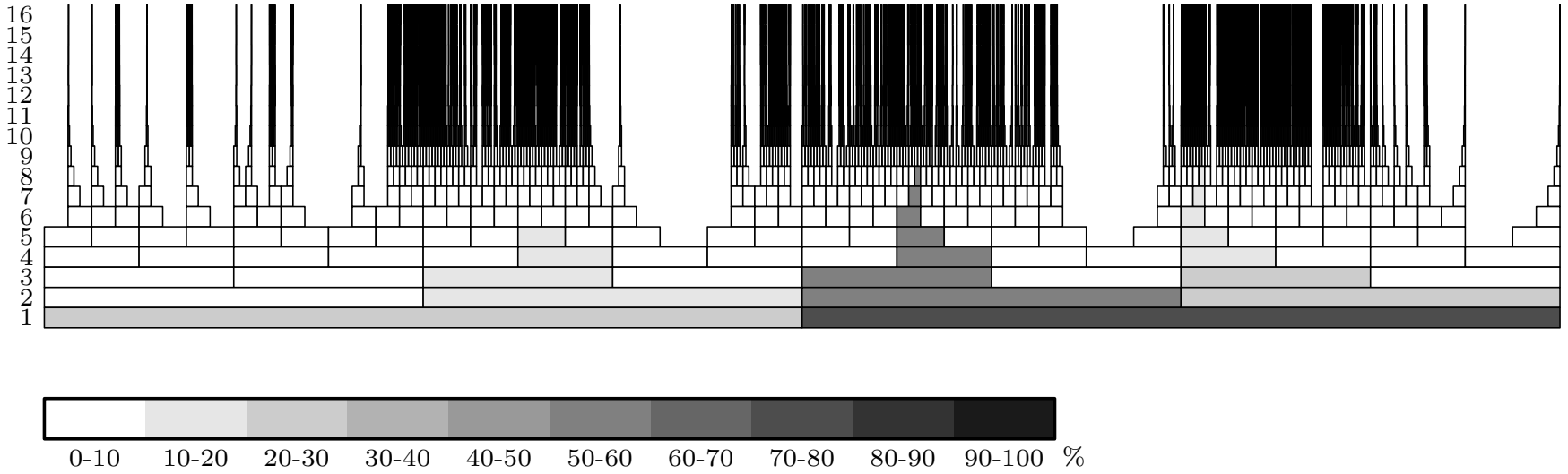
Každý p-agregát obsahuje $2^{32-\mathrm{p}}$ adres. Kratší prefix znamená *větší* množinu adres!

Rozložení adres ukazuje zajímavé struktury na všech úrovních rozlišení (= délka prefixu).

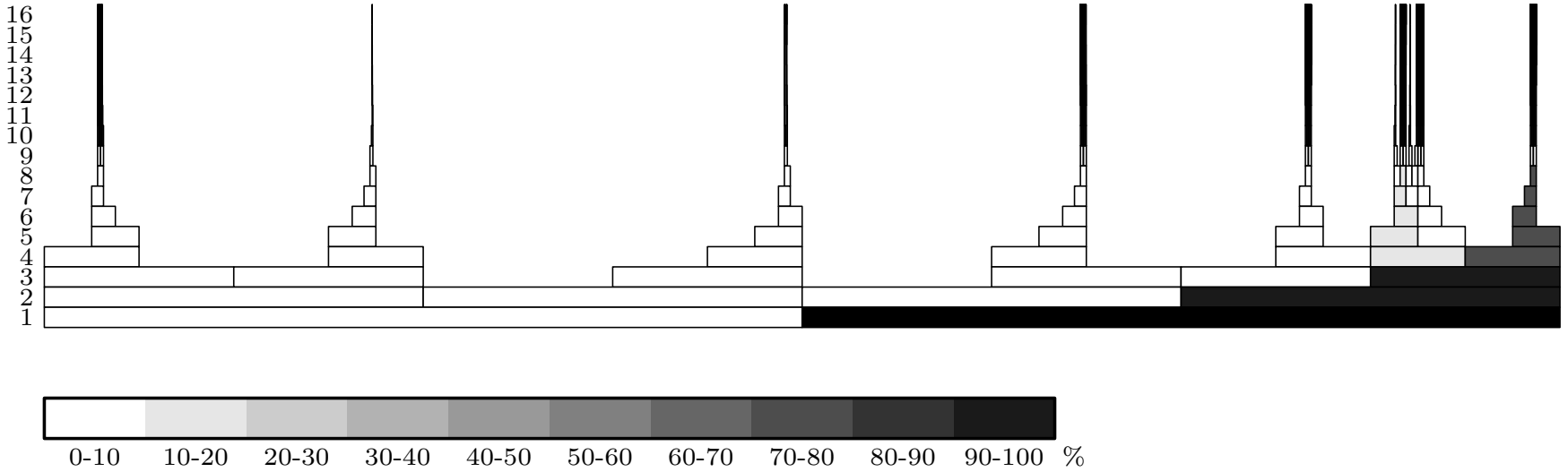


Vzorek provozu (∼4 h.) z přípoje velké univerzity, 62 mil. paketů.

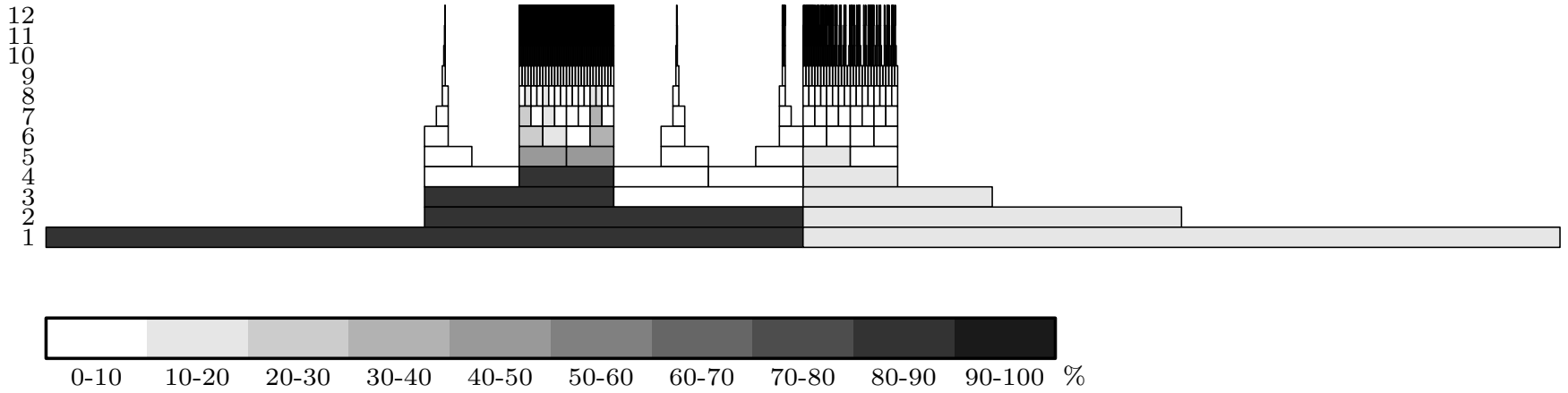
Praha–Brno 10 Gbps



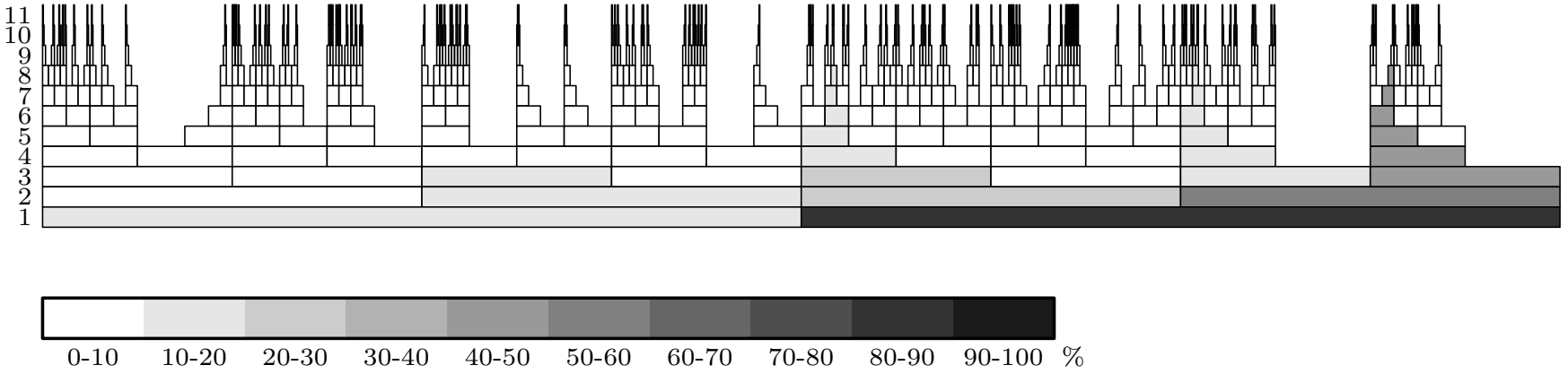
147.0.0.0/8


16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0-10
10-20
20-30
30-40
40-50
50-60
60-70
70-80
80-90
90-100
%

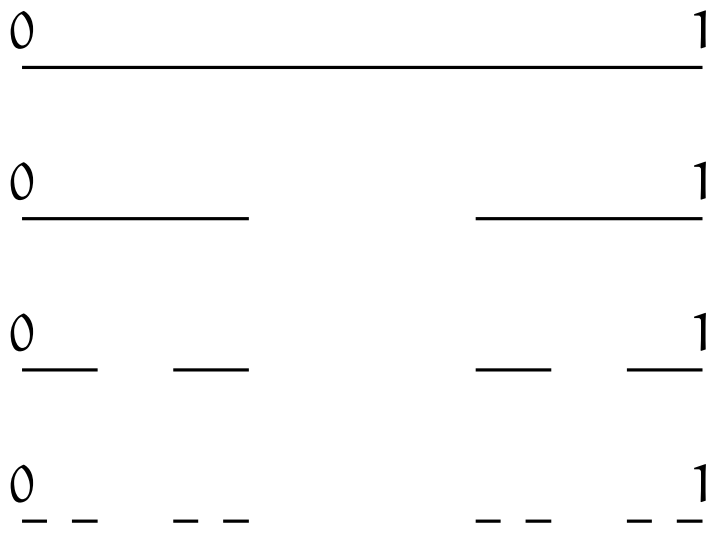
147.224.0.0/12


12
11
10
9
8
7
6
5
4
3
2
1
0-10
10-20
20-30
30-40
40-50
50-60
60-70
70-80
80-90
90-100
%

147.231.204.0/21



Cantorova množina



Z uzavřeného intervalu $[0,1]$ vyjmeme prostřední třetinu, z obou částí $[0,1/3]$ a $[2/3,0]$ znovu prostřední třetinu atd.

Výsledkem limitního procesu této konstrukce je Cantorova množina, která je nekonečná a má dokonce stejnou mohutnost jako celý interval $[0,1]$ ($\aleph$ = nespočetná). Přitom neobsahuje žádný otevřený interval, takže její celková délka (míra) je 0.

Fraktální dimenze

Eukleidovské objekty mají celočíselnou dimenzi – 0 (bod), 1 (úsečka, kružnice, . . .), 2 (obdélník, kruh, . . .). U některých "podivných" objektů ale intuitivně cítíme, že jejich dimenze je někde mezi.

Například Cantorova množina je více než soubor izolovaných bodů, ale méně než úsečka.

Sierpińského síto:

Reálné objekty: pobřežní čára, tepny a žíly v těle, atd.

Výpočet fraktální dimenze

Kolmogorovova kapacita, box-counting dimension

Omezenou množinu A v prostoru dimenze n uzavřeme do n-rozměrné krychle a její stranu prohlásím ze jednotkovou délku.

Tuto krychli rozdělíme na síť krychliček o straně ϵ (jejich celkový počet je tedy $1/\epsilon^n$).

$N_\epsilon(A)$ je počet krychliček obsahujících aspoň jeden bod množiny A.

Zmenšujeme-li ϵ, $N_\epsilon(A)$ roste úměrně s nějakou mocninou ϵ:

$$N_\epsilon(A) \propto \epsilon^{-D(A)}.$$

Číslo $D(A)$ nazveme *fraktální dimenze* množiny A.

$$D(A) = -\lim_{\epsilon \to 0} \frac{\log N_\epsilon(A)}{\log \epsilon} \tag{1}$$

U eukleidovských objektů odpovídá jejich geometrické dimenzi, pro Cantorovu množinu je $\log 2 / \log 3 \approx 0,63$.

Odhad fraktální dimenze z dat

Pro naměřená data nemá limita žádný smysl, můžeme ale sledovat, jak se na nějaké konečné škále chová $\log \mathrm{N}_\epsilon(A)$ proti $\log \epsilon$. Směrnice tečny rozumně blízko 0 udává odhad fraktální dimenze.

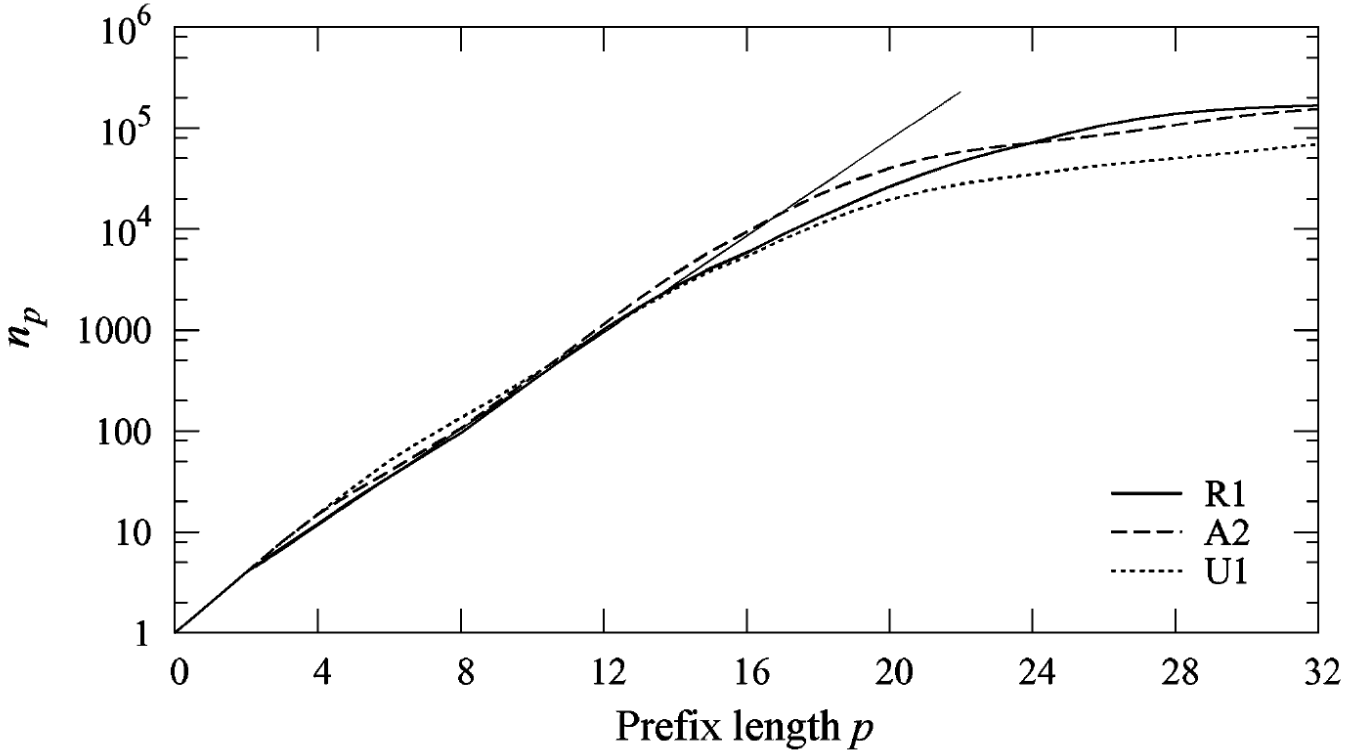
Postup pro IP adresy (A je množina všech zjištěných adres): 32-bitový adresový prostor dělíme postupně na 2, 4, 8, ... části. V p-tém kroku tedy máme (2^{32} je pro nás jednotka délky)

$$\epsilon_\mathrm{p} = 2^{32-\mathrm{p}}/2^{32} = 2^{-\mathrm{p}}$$

Každá "krychlička" v p-tém kroku představuje p-agregát, $\mathrm{N}_{\epsilon_\mathrm{p}}(A) = \mathrm{N}_\mathrm{p}$ je tedy počet prefixů délky p obsahujících aspoň jednu adresu ze vzorku.

$$\mathrm{D}(A) = \lim_{\mathrm{p}\to\infty} \frac{\log \mathrm{N}_\mathrm{p}}{\mathrm{p} \log 2} = \lim_{\mathrm{p}\to\infty} \frac{\log_2 \mathrm{N}_\mathrm{p}}{\mathrm{p}.}$$

Odhad fraktální dimenze vzorků Kohler et al.



Fraktální dimenze vzorku R1 (regionální ISP) je 0,79.

Výpočet v R

```
flows <- read.csv("../4-081103/gn2-flows-10ge.data",
  col.names=c("source", "dest","proto","sip","dip",
  "sport","dport","packets","bytes"))
dip <- flows$dip
lcnt <- c()
for (p in 1:32) {
  lcnt <- c(log(length(unique(dip)), base=2), lcnt)
  dip <- dip %/% 2
}
plot(lcnt)
p <- 5:15
np <- lcnt[p]
fit <- lm(np ~ p)
abline(coef(fit))
```

Multifraktály

Fraktální dimenze je založena na počítání krychliček, které obsahují nějaký bod. Co když ale různé body nemají stejnou váhu?

Míra μ na množině A je zobrazení, které podmnožinám přiřazuje nějaké číslo (váhu). V našich aplikacích to může být např. množinou A adresový prostor a mírou počet paketů příslušejících cílovým adresám dané podmnožiny (např. p-agregátu).

Například ve vzorku GN2

$$\mu(192.8.0.0/16) = 55954.$$

Obvykle nás zajímá, jak se mění míra v okolí nějakého bodu x v závislosti na velikosti tohoto okolí ($\rightarrow$hustota míry). V případě tzv. multifraktálních měr se toto chování často rychle mění podle toho, v jakém bodě x se nacházíme:

$$\mu(B_x(\epsilon)) \propto \epsilon^{\alpha(x)},$$

kde $B_x(\epsilon)$ je okolí bodu x (např. krychlička) o objemu ϵ.

Hölderův exponent

$$\alpha(x) = \lim_{\epsilon \to 0} \frac{\mu(B_x(\epsilon))}{\epsilon}$$

V praktických případech, jako je struktura IP adres, ale nemá limita význam, proto pracujeme s *hrubým* Hölderovým exponentem (coarse-grained H. e.)

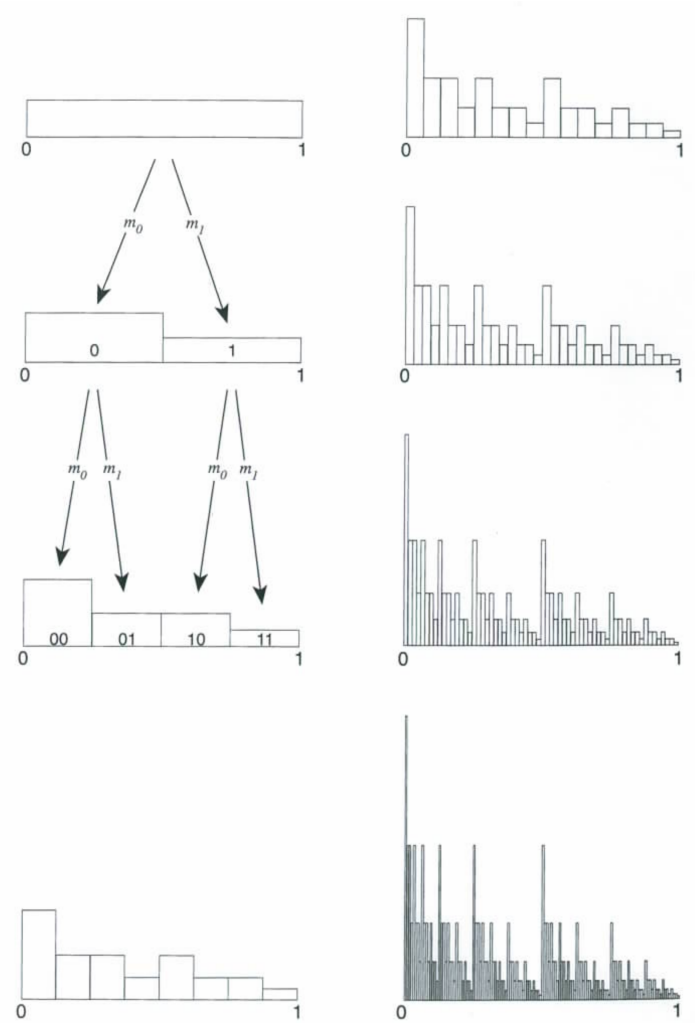
$$\alpha(x) = \frac{\mu(B_x(\epsilon))}{\epsilon}.$$

Chceme nyní posčítat krychličky vždy jen s jednou hodnotou H. e. Ukazuje se, že pro počet výskytů dané hodnoty α platí u multifraktálních měr vztah

$$N_\epsilon(\alpha) \propto \epsilon^{-f(\alpha)}$$

Funkce $f(\alpha)$ nezávisí na ϵ a je zobecněním fraktální dimenze.

Multiplikativní kaskáda



Původně rovnoměrně rozdělená pravděpodobnostní míra na intervalu $[0, 1]$ je rozdělena v poměru $m_0 : m_1$ mezi levou a pravou polovinu, přičemž $m_0 + m_1 = 1$. V každé z těchto polovin je míra rozdělena v poměru $m_0 : m_1$ mezi levou a pravou čtvrtinu atd.

Značení subintervalů v k-tém kroku:

$$I_{0.\beta_0\beta_0\ldots\beta_k}$$

je interval velikosti 2^{-k}, jehož všechny body mají ve dvojkové soustavě stejný začátek uvedený v indexu. Index tedy představuje jeho *levý* krajní bod.

Míry subintervalů

Značíme

$$\mu_{0.\beta_0\beta_0\ldots\beta_k} = \mu(I_{0.\beta_0\beta_0\ldots\beta_k}) = m_0^{n_0} m_1^{n_1},$$

kde n_0 je počet nul a n_1 počet jedniček ve dvojkovém rozvoji čísla $0.\beta_0\beta_0\ldots\beta_k$.

Zajímá nás, jak se s rostoucím k mění závislost míry intervalu vpravo od nějakého bodu $x = 0.\beta_0\beta_0\ldots\beta_k$ v závislosti na jeho délce (2^{-k}):

$$\mu_{0.\beta_0\beta_0\ldots\beta_k} = (2^{-k})^{(\frac{n_0}{k}\nu_0 + \frac{(k-n_0)}{k}\nu_1)},$$

kde $\nu_0 = -\log_2 m_0$ a $\nu_1 = -\log_2 m_1$. Hodnoty ν_0 a ν_0 nezávisejí na k.

Hölderův exponent:

$$\alpha(x) = \frac{n_0}{k}\nu_0 + \frac{(k-n_0)}{k}\nu_1)$$

Předpokládejme, že $m_0 \geq m_1$. Pak platí

$$\nu_0 \leq \alpha(x) \leq \nu_1.$$

Rozložení hodnot Hölderova exponentu

Při daném k, kolik bodů x má H. e. rovný dané hodnotě α?

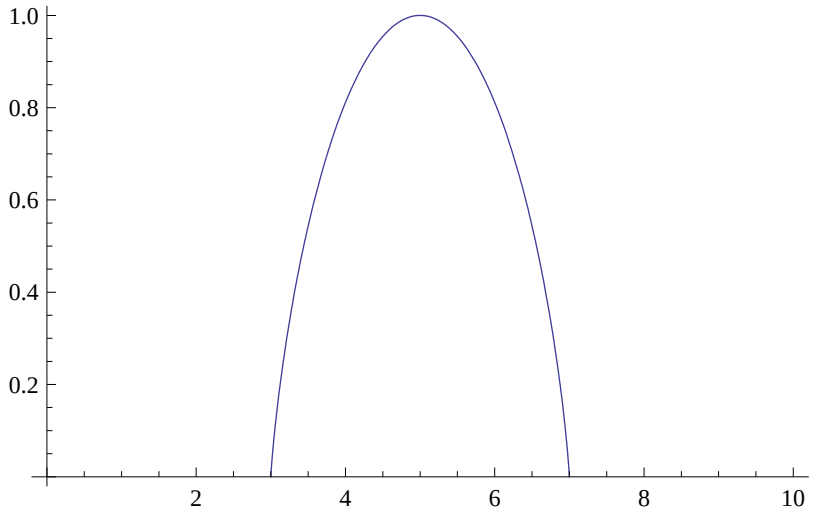
$$N_k(\alpha) = \binom{k}{n_0}$$

Výpočtem se dá ukázat, že tato hodnota aproximovat takto:

$$N_k(\alpha) \sim (2^{-k})^{-f(\alpha)},$$

kde

$$f(\alpha) = -(\frac{\nu_1 - \alpha}{\nu_1 - \nu_0}) \log_2(\frac{\nu_1 - \alpha}{\nu_1 - \nu_0}) - (\frac{\alpha - \nu_0}{\nu_1 - \nu_0}) \log_2(\frac{\alpha - \nu_0}{\nu_1 - \nu_0})$$


1.0
0.8
0.6
0.4
0.2
2
4
6
8
10