Budapesti Műszaki és Gazdaságtudományi Egyetem
Villamosmérnöki és Informatikai Kar

Statisztikai módszerek a skálafüggetlen hálózatok vizsgálatára

Tudományos Diákköri Dolgozat

Gyenge Ádám
adamgyenge@cs.bme.hu

Konzulens: Dr. Telcs András
Számítástudományi és Információelméleti Tanszék

2007

Tartalomjegyzék

1. Bevezetés

Számos műszaki, szociológiai és egyéb tudományos területen figyelhetőek meg olyan jelenségek, melyek Pareto, vagy másnéven hatványfüggvény eloszlású valószínűségi változókkal modellezhetők. Az egyik legfontosabb ilyen terület, az utóbbi években jelentős figyelmet kapó, nagyméretű skálafüggetlen gráfok [1]. Az ilyen gráfok fokszámeloszlása általában Pareto eloszlást követ, és mint kiderült, rengeteg társadalmi-kommunikációs jelenség leírható velük. Ahhoz, hogy ezeket a gráfokat részletesebben tanulmányozni tudjuk, szükségünk van olyan statisztikai módszerekre, melyekkel lehetőségünk nyílik a megfigyelések elemzésére, a vizsgált struktúra „feltérképezésére". Dolgozatomban ezt az igényt igyekszem kielégíteni, bemutatva a felmerülő feladatokat, majd összehasonlítva ezek lehetséges megoldásait szimulált adatokon és valódi adatrendszereken is.

A dolgozat két nagyobb részre tagolható. Elsőként részletesen bemutatom, hogy hogyan illeszthetünk Parero-eloszlást pontok egy halmazára. Ez az eloszlás paramétereinek becslését jelenti a mintaelemek alapján. Általában is az egyik legfontosabb statisztikai kérdés, hogy mik a vizsgált jelenség eloszlásának paraméterei. A skálafüggetlen gráfok esetében kiderült, hogy a fokszámeloszlás paraméterei a gráfok számos tulajdonságát befolyásolják. Több általános, jól használható paraméterbecslő módszer is létezik, de ez a terület jelenleg is aktívan kutatott. Főként az egyes speciális alkalmazásokban jobban teljesítő technikák kerültek előtérbe, ezek hatékonyabbak lehetnek egy-egy adott feladatra.

A 2. fejezetben tehát először áttekintem a klasszikus paraméterbecslő módszereket és alkalmazom őket a Pareto eloszlásra. Ezután bemutatok egy kevésbé ismert módszert is. Ez a feltehetőleg igen általánosan használható lesz, mivel más, abszolút folytonos eloszlásokra is alkalmazható. Itt most a Pareto, illetve kitérőként a normális eloszlásra alkalmazva kerül bemutatásra, és megpróbálom megválaszolni az ezzel kapcsolatban felmerülő kérdéseket. A fejezet végén mérések alapján összehasonlítom a becslési módszereket, szimulációk alapján megpróbálom meghatározni a legjobb becslési eljárást.

A munka második része az illeszkedésvizsgálattal és a skálafüggetlen gráfok fokszámeloszlásának illesztésével foglalkozik részletesen. Itt a fő kérdés, hogy az (esetleg paraméterbecsléssel) felállított modellünkre mennyire illeszkedik a háttérváltozó tényleges eloszlása. Az 3. fejezetben felhasználva a klasszikus χ^2 próbát bemutatunk egy küszöbkereső algoritmust, mely bizonyos keverék eloszlások esetén a Pareto el-

oszlású mintaelemek elkülönítésére képes. A 4. fejetben több gráffejlődési modellt is megvizsgálunk, és megpróbáljuk a fokszámeloszlások és más jellemzők alapján megtalálni ezek közül a valódi hálózatokra legmegfelelőbb eredményt produkáló módszert.

A dolgozatban bemutatott összes módszert és kísérletet megvalósítottam Matlabban. A programok rövid bemutatásával foglalkozik a 5. fejezet. Ezeket és a bemutatott adatrendszereket, méréseket is elérhetővé tettem az alábbi honlapon, így azokat bárki használhatja (az anyagok a dolgozat CD mellékletén is megtalálhatóak):

`http://cs.bme.hu/~adamgyenge`

A 6. fejezetben összefoglalom a dolgozat eredményeit, az A függelékben pedig a matematikai statisztika legfontosabb vonatkozó fogalmai és tételei olvashatóak.

1.1. A Pareto eloszlásról

Az (α, β) paraméterű Pareto eloszlás eloszlásfüggvénye a következőképp van definiálva:

$$F(x)_{\alpha,\beta} = 1 - \left(\frac{x}{\beta}\right)^{-\alpha}, \quad x > \beta \tag{1}$$

Az eloszlás abszolút folytonos, sűrűségfüggvénye:

$$f(x)_{\alpha,\beta} = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{-(\alpha+1)} = \frac{\alpha\beta^{\alpha}}{x^{\alpha+1}}, \quad x > \beta \tag{2}$$

Szükséges, hogy mindkét paraméter pozitív valós szám legyen. A β paraméter az eloszlás „helyét” adja meg, vagyis eltolását a számegyenesen, az α paraméter pedig a lapultságát.

Pareto olasz közgazdász a vagyoni viszonyok megoszlásánál fedezte fel ezt az eloszlást, melyet később róla neveztek el. A tudományos irodalomban azonban számos más név is elterjedt. Sok helyen power-law vagyis hatványfüggvény eloszlásnak nevezik, és ekkor az $\alpha + 1$ exponenst jelölik α-val vagy γ-val. Ekkor tehát a sűrűségfüggvény $f(x) = Cx^{-\alpha}$ (ahol C egy olyan normáló konstans, hogy sürűségfüggvényt kapjunk). Minden itt bemutatott módszer az $\alpha - 1$ visszahelyettesítéssel ekkor is használható. Amennyiben külön nem említjük, mi a (2) jelölést fogjuk használni. Szintén elterjedt még a Zipf-eloszlás elnevezés. Zipf az angol szavak előfordulásainak gyakoriságánál vett észre ilyen eloszlást. A Pareto eloszlásról részletesebben lásd [2]-t.

Az olvasóban joggal merülhet fel az az észrevétel hogy gráfok fokszámai csak diszkrét értékeket vehetnek fel. Ekkor egy pillanatra élve a fenti konvencióval az eloszlás súlyfüggvénye a következő: $p(k) = Ck^{-\alpha}$. Ezt minden k-ra összegezve 1-et kell kapnunk:

$$\sum_k p(k) = C \sum_k k^{-\alpha} = C\zeta(\alpha) = 1$$

Ahol $\zeta(\alpha)$ a Riemann-féle ζ-függvény. Az előbbi egyenletet átrendezve azt kapjuk, hogy $C = 1/\zeta(\alpha)$. Sajnos ennek a függvénynek a számítása igen nehéz, csak numerikusan vagy speciális függvények használatával lehetséges. Ezért paraméterbecslés esetén ezt a diszkrét eloszlást, melyet a legtöbbször Zeta- vagy diszkrét Pareto eloszlásnak neveznek, célszerűbb a folytonos Pareto eloszlással közelíteni, hiszen az értékkészlet olyan naggyá válik, hogy gyakorlatilag folytonosnak tekinthető (vö. [3]).

2. Paraméterbecslés

2.1. Ismert paraméterbecslő technikák

2.1.1. Visszavezetés lineáris regresszióra

A legtöbb kutató ezt a nem túl hatékony módszert használja. Itt elkészítjük a mintaelemek alapértelmezésben egyenletes skálán vett hisztogramját, és kapcsolatot keresünk intervallumoknak megfelelő átlagérték és az intervallumokba eső mintaelemszám között. A empirikus hisztogram abszolút folytonos eloszlás esetén általában jó közelítése az elméleti sűrűségfüggvényeknek, jelen esetben (1)-nek, ha nagy számú intervallumra osztjuk a tartományt és még ennél is jóval nagyobb a mintaelemszám (erről részletesebb lásd [4] 2.1 fejezet).

Jelölje az egyformán h_n hosszúságú, diszjunkt intervallumokat $\Delta_1, ..., \Delta_m$, ν_j pedig legyen a Δ_j-be eső mintaelemek száma ($\sum_j \nu_j = n$). Ekkor

$$f_n^*(x) = \frac{\nu_j}{nh_n}, \qquad x \in \Delta_j$$

a hisztogramfüggvény. Ez közelítése a sűrűségfüggvénynek, így ha az intervallumok középpontjait $x_1, ..., x_m$ jelöli, és $y_i = f_n^*(x_i)$ akkor az $y_i \approx Cx_i^{-(\alpha+1)}$ kapcsolat áll fenn minden i-re, ahol $C = \alpha\beta^\alpha$. Elhagyva az indexeket jól ismert módszer a következő lineáris regresszióra történő visszavezetés:

$$y \approx Cx^{-(\alpha+1)} \Longleftrightarrow \ln y \approx \ln C - (\alpha + 1) \ln x$$

Ha a közelítési hibának a négyzetes hibát tekintjük, akkor a hagyományos lineáris regresszió megoldása adja a legoptimálisabb együtthatókat (ld. [5]):

$$\hat{\alpha} = -\left(1 + \frac{m(\sum_{i=1}^m \ln x_i \ln y_i) - (\sum_{i=1}^m \ln x_i)(\sum_{i=1}^m \ln y_i)}{m(\sum_{i=1}^m (\ln x_i)^2) - (\sum_{i=1}^m \ln x_i)^2}\right) \tag{3}$$

$$\hat{C} = \exp\left(\frac{(\sum_{i=1}^m \ln y_i) + (\hat{\alpha} + 1)(\sum_{i=1}^m \ln x_i)}{m}\right) \tag{4}$$

A $C = \alpha\beta^\alpha$ kifejezést átrendezve azt kapjuk, hogy:

$$\ln \beta = \frac{\ln C - \ln \alpha}{\alpha}$$

ez alapján a becsült β:

$$\hat{\beta} = \exp \frac{\ln \hat{C} - \ln \hat{\alpha}}{\hat{\alpha}} \tag{5}$$

A (3) és az (5) becslésekre a továbbiakban a szakirodalomban elterjedt módon LSF (Least Squares Fitting) becslésként hivatkozom. Heurisztikusan megfogalmazva a változók logaritmusai között kerestünk lineáris kapcsolatot, ami megfelel annak az elterjedt módszernek, hogy az (x, y) párokat log-log skálázású grafikonon ábrázolva egy egyenest illesztenek a pontokra, és ennek meredekségéből következtetnek α-ra. Azonban Goldstein [3] és Newman [2] is felhívják a figyelmet arra, hogy az LSF becslés nagyon pontatlan. Némi javítást lehet elérni például kumulált illesztéssel, logaritmikus intervallumokkal vagy csak az első 5-10 pont figyelembevételével a regressziós együtthatók számításakor. Ezekről részletesebben az említett cikkek írnak. A továbbiakban csak a szimpla regressziót fogom összehasonlítani a többi módszerrel.

2.1.2. Momentum módszer

A momentum módszer esetén kiválasztjuk az eloszlás első k db momentumát (ahol k a paraméterek száma), felírjuk ezeket a paraméterek függvényeiként (pl. $m_j = E(X^j) = g_j(\theta_1, ..., \theta_k)$), majd ha ez a $(g_1, ..., g_k) : R^k \to R^k$ leképezés invertálható (legyen ez az inverz $(h_1, ..., h_k) : R^k \to R^k$), akkor az i-edik paraméterre a $\hat{\theta}_i = h(\hat{m}_1, ..., \hat{m}_k)$ becslést adjuk $(i = 1, ..., k)$, ahol $\hat{m}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$ a minta j-edik empirikus momentuma.

Az (α, β) paraméterű Pareto eloszlás esetében a k-adik momentum a következőképpen számolható ki:

$$m_k = E(X^k) = \int_\beta^\infty x^k \frac{\alpha\beta^\alpha}{x^{\alpha+1}} \mathrm{d}x = \alpha\beta^\alpha \int_\beta^\infty x^{k-\alpha-1} \mathrm{d}x = \alpha\beta^\alpha \left[\frac{x^{k-\alpha}}{k-\alpha}\right]_\beta^\infty$$

Látható, hogy ez utóbbi akkor lesz véges, ha $k < \alpha$. Ekkor a kifejezésre az alábbi érték adódik:

$$m_k = \frac{\alpha\beta^\alpha}{k-\alpha}(0 - \beta^{k-\alpha}) = \frac{\alpha\beta^k}{\alpha - k} \tag{6}$$

Mivel két paramétert keresünk ezért az első két momentumra lesz szükségünk. Azonban ezek csak akkor léteznek, ha $\alpha > 2$. Ekkor

$$m_1 = \frac{\alpha\beta}{\alpha - 1} \tag{7}$$

és

$$m_2 = \frac{\alpha\beta^2}{\alpha - 2} \tag{8}$$

Ez az $(\alpha, \beta) \to (m_1, m_2)$ leképezés akkor és csak akkor invertálható, ha a Jacobi determinánsa nem 0:

$$J = \begin{vmatrix} \frac{\beta(\alpha-1)-\alpha\beta}{(\alpha-1)^2} & \frac{\alpha}{\alpha-1} \\ \frac{\beta^2(\alpha-2)-\alpha\beta^2}{(\alpha-2)^2} & \frac{2\alpha\beta}{\alpha-2} \end{vmatrix} =$$

$$= \frac{(\beta\alpha-\beta-\alpha\beta)}{(\alpha-1)^2}\frac{2\alpha\beta}{\alpha-2} - \frac{\alpha}{\alpha-1}\frac{(\beta^2\alpha-2\beta^2-\alpha\beta^2)}{(\alpha-2)^2} =$$

$$= \frac{2\alpha\beta^2}{(\alpha-2)^2(\alpha-1)} - \frac{2\alpha\beta^2}{(\alpha-1)^2(\alpha-2)} =$$

$$= \frac{2\alpha\beta^2((\alpha-1)-(\alpha-2))}{(\alpha-1)^2(\alpha-2)^2} = \frac{2\alpha\beta^2}{(\alpha-1)^2(\alpha-2)^2}$$

Mivel mindkét paraméter szigorúan pozitív, ezért mind a nevező mind a számláló is az, és így a determináns semmiképpen sem lehet 0, így a leképezés invertálható.

(7) alapján:

$$\beta = \frac{m_1(\alpha-1)}{\alpha} \tag{9}$$

ezt (8)-ba behelyettesítve:

$$m_2 = \frac{\alpha}{\alpha-2}m_1^2\frac{(\alpha-1)^2}{\alpha^2} = \frac{\alpha^2-2\alpha+1}{\alpha^2-2\alpha}m_1^2$$

Ebből a következő másodfokú egyenlet adódik α-ra:

$$(m_2-m_1^2)\alpha^2 - 2(m_2-m_1^2)\alpha - m_1^2 = 0$$

Az egyenlet megoldásai:

$$\alpha_{1,2} = \frac{2(m_2-m_1^2) \pm \sqrt{4(m_2-m_1^2)^2 + 4(m_2-m_1^2)m_1^2}}{2(m_2-m_1^2)} =$$

$$= 1 \pm \sqrt{1 + \frac{m_1^2}{m_2-m_1^2}}$$

Könnyű látni, hogy $m_1^2 < m_2$, és így a gyök alatt mindig egy 1-nél nagyobb szám fog állni. Ennek a négyzetgyöke is nagyobb 1-nél, így mindig csak az $\alpha = 1 + \sqrt{1 + m_1^2/(m_2-m_1^2)}$ megoldás lesz pozitív.

Mivel $m_2 - m_1^2$ tulajdonképpen a szórásnégyzet, ezért ezt az empirikus szórásnégyzettel, m_1-t pedig a mintaátlaggal helyettesítve az alábbi becslést adhatjuk α-ra:

$$\hat{\alpha} = 1 + \sqrt{1 + \frac{\bar{x}^2}{s_n^2}} \tag{10}$$

β-ra pedig (9) alapján:

$$\hat{\beta} = \frac{\hat{\alpha} - 1}{\hat{\alpha}} \bar{x} = \frac{\sqrt{1 + \frac{\bar{x}^2}{s_n^2}}}{1 + \sqrt{1 + \frac{\bar{x}^2}{s_n^2}}} \bar{x} \tag{11}$$

A (10) és (11) becslőkre az összehasonlítások során MoE-ként (Moments Estimator) hivatkozom.

2.1.3. Maximum likelihood módszer

A statisztikában a paraméterbecslés egyik legelterjedtebb módja a maximum likelihood módszer [4]. Ennek lényege, hogy egy adott kimenetel esetén keressük azt a paramétert, amire annak a realizációnak a valószínűsége a legnagyobb. Jelölje az n elemű minta realizációját $x_1, ..., x_n$. Ekkor a likelihood egyenlet:

$$L_{\alpha,\beta}(\mathbf{x}) = \prod_{i=1}^{n} I(x_i > \beta) \frac{\alpha}{\beta} \left(\frac{x_i}{\beta} \right)^{-(\alpha+1)} = I(x_1^* > \beta) \prod_{i=1}^{n} \frac{\alpha \beta^\alpha}{x_i^{\alpha+1}} \tag{12}$$

Ezt a kifejezést szeretnénk maximalizálni α-ban és β-ban. Látható, hogy a számlálóban β^α van. Jó tehát, ha β minél magasabb, így (12) is nagyobb lesz. Nem lehet viszont magasabb x_1^*-nál, vagyis a rendezett minta első eleménél. Így az ML becslés β-ra:

$$\hat{\beta} = x_1^* = \min x_i \tag{13}$$

A likelihood egyenlet logaritmusát véve a log-likelihood egyenletet kapjuk, melynek ugyanott van a maximum helye, mint az eredeti likelihood egyenletnek:

$$l_{\alpha,\beta}(\mathbf{x}) = \ln \prod_{i=1}^{n} \frac{\alpha x_1^{*\alpha}}{x_i^{\alpha+1}} \cdot I(x_i > x_1^*) =$$

$$= \sum_{i=1}^{n} \ln \alpha + \sum_{i=1}^{n} \alpha \ln x_1^* - \sum_{i=1}^{n} (\alpha + 1) \ln x_i$$

Ennek az α szerinti a szélsőértékét deriválással kereshetjük meg:

$$\frac{\partial l_{\alpha,\beta}(\mathbf{x})}{\partial \alpha} = \sum_{i=1}^{n} \frac{1}{\alpha} + \sum_{i=1}^{n} \ln x_1^* - \sum_{i=1}^{n} \ln x_i = \frac{n}{\alpha} + n \ln x_1^* - \sum_{i=1}^{n} \ln x_i$$

Ezt a kifejezést 0-ra megoldva a következő becslést kapjuk, melyről könnyen látható, hogy valóban maximumhelye lesz a likelihood egyenletnek:

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n} \ln x_i - n \ln x_1^*} \tag{14}$$

Malik [6] megmutatta, hogy ezek a becslések torzítottak. Saksena és Johnson [7] az ML becslés függvényeiként levezették az elégséges, teljes és torzítatlan becslést a Pareto eloszlás paramétereire:

$$\hat{\alpha} = \left(\frac{n-2}{n}\right) \cdot \hat{\alpha}_{MLE} = \frac{n-2}{\sum_{i=1}^{n} \ln x_i - n \ln x_1^*} \tag{15}$$

$$\hat{\beta} = \left(1 - \frac{1}{(n-1)\hat{\alpha}_{MLE}}\right) \cdot \hat{\beta}_{MLE} = \left(1 - \frac{\sum_{i=1}^{n} \ln x_i - n \ln x_1^*}{n(n-1)}\right) \cdot x_1^* \tag{16}$$

Ezek a Lehmann-Scheffé-tétel miatt hatásosak (Lásd A függelék), vagyis legkisebb szórású becslők az összes torzítatlan becslés közül.

Mostantól (13) és (14) becslőket MLE, a (15) és (16) becslőket pedig CMLE, vagyis korrigált ML becsléseknek nevezzük.

Megjegyzés: Egy további nevezetes becslési módszer a Bayes-becslés, mely négyzetes rizikó értelemben az optimális eredményt adja. Ehhez azonban szükséges ismerni a paraméterek *a priori* eloszlását. Ennek hiányában egyenletes eloszlás tételezhető fel a teljes paramétertérben. Ez most azonban nem lehetséges, mert a paramétertér egy teljes síknegyed (ahol mind α mind β pozitív), ami nem korlátos, és így rajta az egyenletes eloszlás nem értelmezhető. Ha viszont ennek egy korlátos tartományára szorítkoznánk, akkor ezen kívüli valódi paraméterek esetében használhatatlan lenne a becslésünk.

2.2. A csonkított momentumok módszere

2.2.1. Elméleti megfontolások

A módszer W. Glänzel alábbi tételen alapul [9], melynek G. G. Hamedani által leírt verzióját közöljük [10]:

1. Tétel (Glänzel). *Legyen (Ω, A, P) egy valószínűségi mező és legyen $H = [a, b]$ egy intervallum valamilyen $a < b$-re ($a = -\infty$ és $b = +\infty$ is megengedett). Legyen $X : \Omega \to H$ egy folytonos valószínűségi változó F eloszlásfüggvénnyel és legyen g és h két H-n értelmezett valós értékű függvény úgy, hogy*

$$E\{g(X)|X \geq x\} = E\{h(X)|X \geq x\}\lambda_h^g(x), \quad x \in H$$

értelmezett valamilyen λ_h^g valós függvényre. Tegyük fel, hogy $g, h \in C^1(H)$, $\lambda_h^g \in C^2(H)$ és F kétszer folytonosan differenciálható és szigorúan monoton függvény a H halmazon. Végül, tegyük fel, hogy a $h\lambda_h^g = g$ egyenletnek nincs megoldása H belső pontjaiban. Ekkor g, h, és λ_h^g egyértelműen meghatározzák F-et, nevezetesen

$$F(x) = \int_a^x C \left| \frac{\lambda'(u)}{\lambda(u)h(u) - g(u)} \right| \exp(-s(u))du$$

ahol az s függvény az $s' = \frac{\lambda' h}{\lambda h - g}$ differenciálegyenlet megoldása és C egy konstans úgy megválasztva, hogy $\int_H dF = 1$ legyen.

A tétel heurisztikus jelentése az, hogy minden abszolút folytonos eloszlású X valószínűségi változóra megfelelő g, h és λ függvényekkel igaz, hogy:

$$E(g(X)|X \geq x) = E(h(X)|X \geq x)\lambda(x) \tag{17}$$

továbbá ezek a függvények egyértelműen jellemzik az adott eloszlást.

Ez az összefüggés a valószínégi változókból csonkítással kapott újabb váltózok bizonyos transzformáltjainak várható értéke között állít fel kapcsolatot. A csonkítás statisztikai szempontból azt jelenti, hogy egy mintából csak az x-nél nagyobb elemeket fogadjuk el érvényesnek, a többi elemről megfeledkezünk.

[11]-ben a szerzők a legtöbb nevezetes eloszlásra megadják ezeket a függvényeket, így a Pareto eloszlást jellemzőket is. Ezek ebben az esetben a következők:

$$g(x) \equiv 1 \qquad h(x) = \frac{\alpha + 1}{x} \qquad \lambda(x) = \frac{x}{\alpha}$$

Így tehát (17)-et átrendezve, és ezen függvényeket behelyettesítve a következő kifejezésre jutunk, ha X egy Pareto eloszlású valószínűségi változó:

$$1 - E((\alpha + 1)X^{-1}|X \geq x)\frac{x}{\alpha} = 1 - \frac{\alpha + 1}{\alpha}E(X^{-1}|X \geq x)x = 0 \tag{18}$$

Erre egy másik bizonyítást is adhatunk az alábbi érdekes állítást felhasználva:

1. Állítás. *Ha X Pareto eloszlású (α, β) paraméterekkel, akkor $Y_x = X|X \geq x$ Pareto eloszlású (α, x) paraméterekkel minden rögzített $x > \beta$ esetén.*

Bizonyítás: Legyen X eloszlásfüggvénye $F_{\alpha,\beta}(x) = P(X < x) = 1 - \left(\frac{x}{\beta}\right)^{-\alpha}$

Y_x eloszlásfüggvénye:

$$G(y) = P(X < y|X > x) = \frac{P(X < y, X > x)}{P(X > x)} = \frac{F(y) - F(x)}{1 - F(x)} =$$

$$= \frac{1 - (y/\beta)^{-\alpha} - 1 + (x/\beta)^{-\alpha}}{(x/\beta)^{-\alpha}} = 1 - \left(\frac{y}{x}\right)^{-\alpha}, \quad y > x$$

ez pedig pont az (α, x) paraméterű Pareto eloszlás eloszlásfüggvénye. □

Az 1. állítást és a momentumokra vonatkozó (6) képletet felhasználva a csonkított valószínűségi változó tetszőleges momentuma ($k < \alpha$ esetén):

$$E(X^k|X > x) = \frac{\alpha x^k}{\alpha - k} \tag{19}$$

Ebbe $k = -1$-et helyettesítve, és átrendezve kapjuk a (18)-as egyenletet.

2.2.2. Gyakorlati megvalósítás

Az 1. tétel alapján egy új módszert adhatunk az α paraméter becslésére. (18)-at most felhasználva, keressük azt az α-t amire a

$$E\left(1 - \frac{\alpha + 1}{\alpha} E(X^{-1}|X \geq x)x\right)^2 \tag{13}$$

kifejezés minimális, vagyis a hibát négyzetes értelemben minimalizáljuk.

Jelölje a minta realizációját $x_1, ..., x_n$. Ennek alapján egy t pontban megkaphatjuk a -1-edik csonkított momentum függvény értékének becslését, ha elhagyjuk a t-nél kisebb mintaelemeket, és a megmaradóak reciprokának vesszük az átlagát:

$$\hat{v}(t) = \frac{1}{\#\{x_i|x_i \geq t\}} \sum_{x_i \geq t} x_i^{-1}$$

Ezt megcsinálhatjuk több t értékre is, jelölje ezeket a pontokat $t_1, ..., t_m$. Ezeket megfelelően megválasztva a $v(t) = E(X^{-1}|X \geq t)$ függvényt közelítő pontokat kapunk. De felmerül a kérdés, hogy hogyan válasszuk meg ezeket a csonkítási pontokat, vagyis mely pontokba becsüljük a $v(t)$ függvényt? Erre több válasz is adható, itt most két természetesen adódó lehetőséget említünk:

1. Adott számú (pl. 10, 100,... db) pont egyenletesen eloszlatva a mintaterjedelem által kijelölt intervallumban.

2. A csonkítási pontok legyenek maguk a mintaelemek ($t_i = x_i$), vagy általuk meghatározott pontok, pl. minden egymást követő mintaelempár által kijelölt intervallum felezőpontja ($t_i = (x_i + x_{i+1})/2$).

A két különböző esetet az 1. ábra[1] mutatja be. Mindkét esetben igaz az, hogy míg a mintaterjedelem alsó végénél csonkítva még viszonylag sok mintaelem marad az átlagoláshoz, addig a felső végéhez közeledve már egyre kevesebb. Ebből következően ezek a pontok már jóval pontatlanabbul közelítik a $v(t)$ függvény valódi értékeit. Az ábrákon is látszik és mérések alapján is kiderült, hogy a csonkítási pontokat az 2. módszer szerint célszerű megválasztani, a csonkítási pontokat az x_i értékekhez vesszük fel.

Így tehát a csonkított várható érték képzés egy transzformáció a mintaelemeken, az $\underline{x}$ vektorból kapunk egy $\underline{v}$ vektort. Ezt szorozzuk meg elemenként a csonkítási pontokat tartalmazó $\underline{t}$ vektorral (ami jelen esetben $\underline{x}$), az ereményt jelöljük $\underline{e}_{-1,1}$-el (ui. a -1-edik momentumot szorozzuk x első hatványával). Tehát

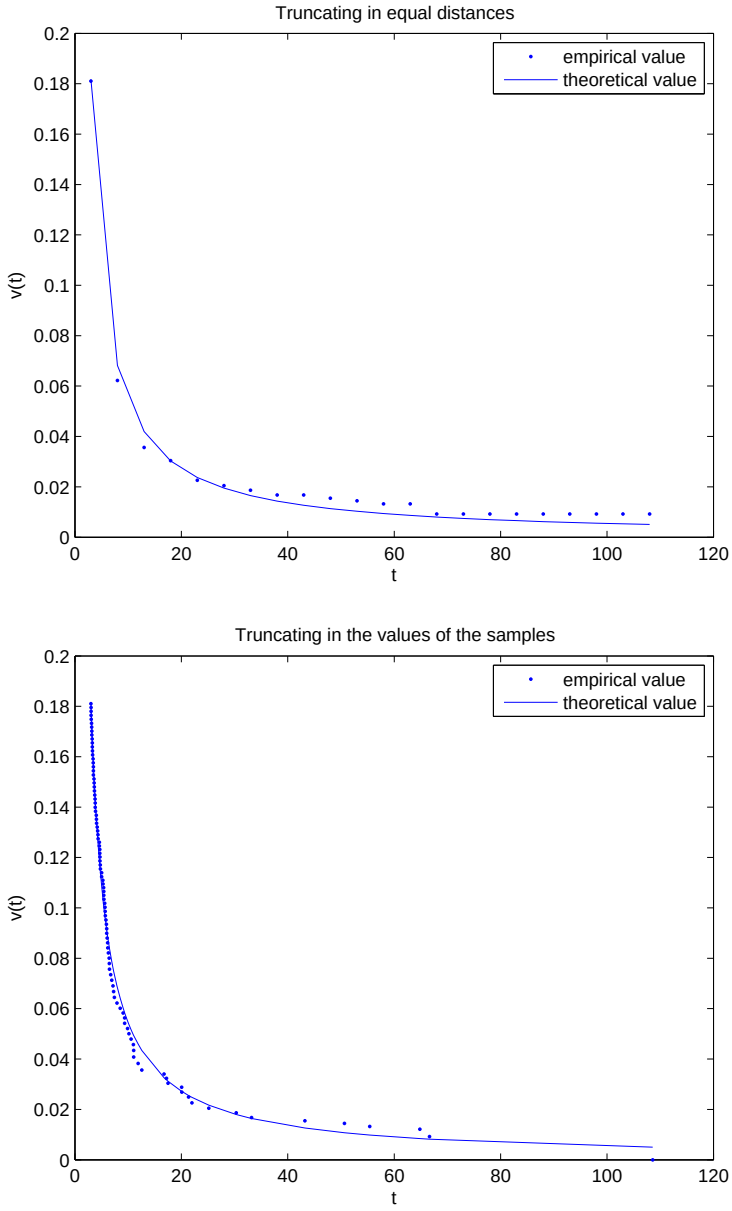
$$\underline{e}_{-1,1} = \left(\frac{x_1}{\#\{x_i | x_i \geq x_1\}} \sum_{x_i \geq x_1} x_i^{-1}, \frac{x_2}{\#\{x_i | x_i \geq x_2\}} \sum_{x_i \geq x_2} x_i^{-1}, ... \right)$$

Vezessük be továbbá az $a = \frac{\alpha+1}{\alpha}$ jelölést. Ekkor közelítőleg az $1 - a \cdot \underline{e}_{-1,1} \approx 0$ kapcsolat áll fenn az új vektorra. Erre alkalmazhatjuk a már említett hagyományos lineáris regressziót, és az annak eredményeként kapott $\hat{a}$ együtthatóval értelemszerűen a következő becslést adhatjuk α-ra:

$$\hat{\alpha} = \frac{1}{\hat{a} - 1} \tag{14}$$

Ez a módszer azért fog jobb eredményt adni, mint a sima lineáris regresszió, mert a teljes mintában lévő információ megjelenik v_1-ben, egy elemet kivéve a maradékban levők v_2-ben, stb. Így ezek sokkal pontosabb becslései a $v(t)$ függvénynek, mint a hisztogram a sűrűségfüggvénynek, és így az $e_{-1,1}$ is sokkal simább lesz. Természetesen ez már kevésbé igaz a nagy értékekre, ahol már csak kevesebb pont számít bele az átlagolásba (lásd később).

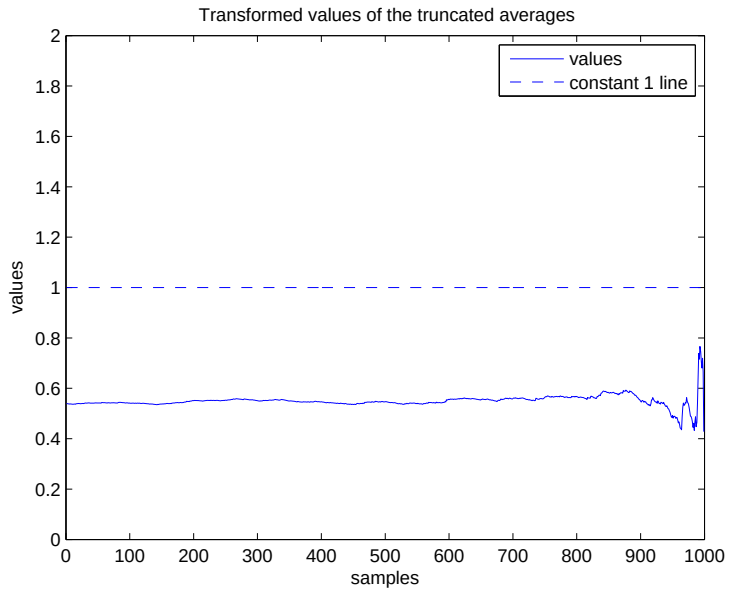
[1] **Megjegyzés:** Ez és a többi ábra is angol nyelvű programmal készült



1. ábra. A $v(t)$ függvény értékei egy 100 elemű minta ($\alpha = 1.2$, $\beta = 3.0$) alapján különböző csonkítási pontokat választva (a. egyenlő távolságra lévő pontok, b. a mintaelemek)

Mivel az egyenletekben β nem szerepel, ezért ezzel a módszerrel rá nem lehet becslést adni. Ez rögtön meg is mutatja a módszer egyik gyenge pontját. Ennél és más eloszlásoknál is csak azok a paraméterek becsülhetőek, amelyek szerepelnek az egyenletekben, és így karakterizálják az eloszlást. Elsőre furcsának hangzik, hogy egy paraméter ilyen lehet, de pont a Pareto eloszlás példája mutatja, hogy lehet ilyen. További hasonló esetek lehetségesek, amikor a paraméterek az eloszlás tartóját befolyásolják és nem a lecsengését, alakját, stb.

Ha a csonkítási pontokat a mintaelemek helyére választjuk, még mindig jelentős szórást tapasztalhatunk a magasabb rangú pontoknál. Ezt szemlélteti a 2. ábra.



2. ábra. Egy 1000 elemű mintából ($\alpha = 1.2$, $\beta = 3.0$) az $E(X > x_i) \cdot x_i$ transzformációval nyert értékek a csonkítási pont rendezett mintabeli rangja alapján megjelenítve (egyenes vonal) és a konstans 1 egyenes. A két egyenes közötti regressziós együttható közelítőleg $\frac{\alpha}{\alpha+1}$.

Látható, hogy a rendezett mintaelemek által alkotott görbe (amelyhez szeretnénk megfelelő együtthatót találni, amivel megszorozva minél jobban közelíti a konstans 1 függvényt) a kisebb rangú pontoknál még szinte teljesen egyenes, a felső negyedben azonban már jócskán fluktuál. Ez több okra is visszavezethető. Egyrészt például a mintaelemekből készített hisztogramfüggvény is jelentős mértékben fluktuál ebben

a tartományban (erről részletesebben [2] ír). Másrészt a már említett jelenség is azt okozza, hogy itt jóval kevesebb pont „marad életben" a csonkítás után, és így jóval pontatlanabbul tudjuk csak közelíteni a várható értéket. A fluktuáció kiküszöbölésére az javasolható, hogy a pontoknak legfeljebb csak az alsó 50-70%-át vegyük figyelembe az illesztéskor, így jóval pontosabb eredmények születhetnek.

A módszer hatékony implementációjakor, például nagyméretű mintákon végrehajtott adatbányászati felhasználásokban gondot jelenthet a csonkított átlagok kiszámítása. Erre dinamikus programozást érdemes alkalmazni, ha a mintát már rendeztük. Ha n db mintánk van, töltsük fel egy n méretű tömb elemeit csökkenő sorrendben, ahol az i-edik cella $E(X|X > x_i)$ becslését tartalmazza. A kitöltés közben tartsuk nyílván, hogy mennyi az aktuális x_i pontnál nagyobb értékű elemek reciprokainak összege és darabszáma, kezdetben mindkettő 0. A kitöltés közben vegyük a sorban eggyel kisebb mintaelemet és adjuk hozzá az összeghez a reciprokát, az össz-számhoz pedig 1-et, ezek alapján az átlag egy osztással nyerhető, ezt szorozzuk meg az aktuális mintaelemmel. Általános esetben ha k különböző momentumot kell kiszámolnunk, akkor egy $k \times n$-es mátrixot kell értelemszerűen kitölteni. Közben már a regressziós együtthatók számításának első lépését (az átlagolást) is elvégezhetjük, így még egy végigolvasást megspórolhatunk.

Az ebben a fejezetben bemutatott becslési módszert TM becslőnek (Truncated Moments Estimator) nevezem a továbbiakban.

2.2.3. Egy másik példa: a normális eloszlás

A természet- illetve társadalomtudományokban az egyik leggyakrabban előforduló eloszlás a Gauss-, avagy normális eloszlás. Ezért most bemutatjuk a csonkított momentumok módszerét erre az eloszlásra is. Az (m, σ) paraméterek esetén a következő egyenletek jellemzik a csonkított momentumokat [11]:

$$g(x) = x^2 - mx - \sigma^2 \qquad h(x) = x - m \qquad \lambda(x) = x$$

Ezeket (10)-be behelyettesítve a következő összefüggés adódik:

$$E(X^2 - mX - \sigma^2 | X \geq x) - E(X - m | X \geq x)x = 0 \tag{20}$$

A várhatóérték linearitásából következően így

$$E(X^2|X \geq x) - mE(X|X \geq x) - \sigma^2 - E(X|X \geq x)x + mx = 0$$

Ezt a már bemutatott jelölésrendszerre átírva:

$$(e_{2,0} - e_{1,1}) - m(e_{1,0} - e_{0,1}) - \sigma^2 = 0 \tag{21}$$

Ebben pedig a szokásos egyváltozós lineáris regresszióval meghatározhatóak az m és a σ paraméterek

2.3. A becslési módszerek összehasonlítása

A bemutatott becslési eljárásokat többféle szerint össze lehet hasonlítani. Statisztikai nézőpontból a legfontosabb kérdések a becslés torzítása (a várható értéke mennyire tér el a becsült paramétertől) és torzítatlanság esetén a hatásossága (mekkora szórással becsli a paramétert). Számítási hatékonyság szempontjából nem elhanyagolható az algoritmusok költsége sem, különösen nagy minták esetén. A módszereket szimuláció segítségével hasonlítottuk össze: nagyméretű mintákat generáltunk, és ezeken teszteltük a módszerek hatékonyságát.

Leginkább az α paraméter becslésére voltunk kíváncsiak és a TM becsléssel csak ez becsülhető, ezért elsőként erre hasonlítottuk össze a módszereket. Erre vonatkozó méréseinket 10000 méretű, $\alpha = 2.5$, $\beta = 3$ paraméterű mintákra az 1. táblázat tartalmazza. Itt most $\alpha > 2$, így a MoE is használható.

Becslési módszer	átlag	szórás
MLE	2.5005	0.024
CMLE	2.5000	0.024
MoE	2.6001	0.119
LSF	1.8992	0.257
TME	2.5004	0.027

1. táblázat. Statisztikai eredmények az α becslésére 10000 méretű, α=2.5, β=3 paraméterű minta esetén, 50 futás eredményeit átlagolva

Az 1. táblázatban látható eredmények összhangban vannak Goldstein és társainak [3]-beli eredményeivel. Ők a fentiek közül csak az MLE és LSF becsléseket hasonlították össze, és ezek közül az MLE bizonyult a legjobbnak. Látható, hogy viszonylag torzítatlan becslést csak az MLE, CMLE és a TME ad, ezekre az eltérés

kisebb mint 0.01%. Közülük is az elméleti eredményeknek megfelelően a CMLE a legpontosabb, ez adja a legjobb eredmény mellett a legkisebb szórást is, vagyis a leghatásabb. Ezzel szemben a lineáris regressziós módszer meglepően nagy torzítású és szórású. A momentum-módszerrel kapott becslés is túl pontatlan. A TME viszont ugyan torzítatlan, de szórása kicsit nagyobb mint a CML becslésé.

A β becslések összesonlítása hasonló módon történt, itt a TM becslés már nem volt használható. Szintén egy 10000 méretű mintát generáltunk, és azon hasonlítottuk össze a módszereket.

Becslési módszer	átlag	szórás
MLE	3.0003	0.0003
CMLE	3.0000	0.0003
MoE	89.3576	210.72
LSF	1039	3.065

2. táblázat. Statisztikai eredmények a β becslésére 10000 méretű, α=2.5, β=3 paraméterű minta esetén, 50 futás eredményeit átlagolva

A mérések eredményei a 2. táblázatban találhatóak. Ebben a feladatban az MLE, és méginkább a CMLE páratlanul erősnek bizonyult. A CMLE-vel pontosan és nagyon kicsi szórással lehetett meghatározni a β-t. A MoE és az LSF módszerek viszont kiábrándítóan rosszul becsültek, az eredményeknek nem sok közük volt a valósághoz. Ennek a jelenségnek az okára egyenlőre nem találtunk magyarázatot.

Habár a CMLE a Pareto eloszlás esetén jobbnak bizonyul, mint a TME, nem szabad azonban lebecsülnünk e módszer jelentőségét. A Maximum Likelihood egyenletet ugyanis sok esetben csak numerikusan lehet megoldani. A TME becslés olyan eloszlásokra, melyek karakterizáló egyenletei lineáris regresszióra visszavezethetőek (pl. Normális, Gamma, Beta) könnyen alkalmazható az itt bemutatotthoz hasonló módon. Ezek az egyenletek nevezetes eloszlások hatalmas gyűjteményére rendelkezésre állnak, másrészt új eloszlások esetén is az 1. tétel differenciálegyenleteinek megoldásaival megtalálhatóak. Nem-lineáris esetben általánosított regresszó (pl. ACE algoritmus, ld. [4] 8.4 fejezet) is használható, azonban ennek a módszernek a felderítésére további kutatás szükséges.

Összefoglalásként elmondható, hogy a Pareto eloszlás paramétereinek becslésénél

a CMLE-t érdemes használni, nemcsak a pontos és precíz eredmények, de a könnyű implementálhatóság és gyors futás miatt is. Ez a módszer minden szempontból a legoptimálisabb eredményt adja.

3. Illeszkedésvizsgálat

3.1. Illeszkedésvizsgálat a Pareto eloszlásra

A paraméterbecslés mellett egy másik fontos statisztikai kérdés, hogy a vizsgált jelenség mennyire illeszkedik egy adott eloszlásra, modellre. Ezt illeszkedésvizsgálatnak nevezik, mely a hipotézisvizsgálat egy speciális esete. A null-hipotézisünk jelen esetben az, hogy a vizsgált változó Pareto eloszlást követ, ezt vetjük össze azzal az ellenhipotézissel, hogy nem. Több elterjedt próba is van ennek eldöntésére, a két legismertebb ezek közül a χ^2 próba és a Kolmogorov-Smirnov próba (mindkét próba részletes leírása megtalálható [4]-ben). Több kutató is elköveti azt a hibát, hogy KS tesztet alkalmaz a skálafüggetlen gráfok esetén (lásd pl. [3]). Azonban a KS teszt kizárólag folytonos esetben használható. Habár a paraméterbecslés esetén a fokszám-eloszlást egy folytonos eloszlással közelítettük, ezt a KS próbánál nem tehetjük meg, ugyanis a próbastatisztika határeloszlása egyszerűen nem az lesz, mintha valóban folytonos lenne a háttérváltozó.

Ebből kifolyólag a KS próba ebben az esetben nem használható. Léteznek újabb illeszkedésvizsgálati módszerek is, azonban ezeknek általában nehézkes a számításuk. Ezért most a klasszikus χ^2 próbát használjuk. Ennek alapja, hogy tetszőleges $A_1, ..., A_r$ teljes eseményrendszer esetén, $\nu_1, ..., \nu_r$-el jelölve az egyes események abszolút gyakoriságát ($\sum_{i=1}^{r} \nu_i = n$), továbbá a H_0 null-hipotézis fennállása és $n \to \infty$ esetén a

$$\sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i} \tag{22}$$

próbastatisztika eloszlásban tart az $(r-1)$-ed fokú χ^2 eloszláshoz, ahol p_i jelöli az A_i esemény valószínűségét H_0 fennállása esetén.

Statisztikai próbáknál meghatározzuk a próbastatisztika értékét a minták alapján, és ezt hasonlítjuk össze a határeloszlás $1-\varepsilon$ kvantilisével. Amennyiben a statisztika értéke nagyobb a megfelelő kvantilisnél, úgy elutasítjuk a null-hipotézist, ellenkező esetben elfogadjuk [2]. Az ε érték lesz az elsőfajú hiba valószínűsége, vagyis annak az esélye, hogy a null-hipotézis igaz, de mégis elutasítjuk. Az illeszkedés jóságának mértéke a szignifikancia (bizonyosság). Ez azt jelenti, hogy milyen biztonsággal utasítjuk/fogadjuk el a null-hipotézist. Ez tulajdonképpen az a legmagasabb ε ami

[2] Diszkrét esetben még előfordulhat randomizálási tartomány is, amikor egy véletlen szám alapján döntünk az elfogadásról. Részletesebben lásd [4] 4.1 fejezete

mellett még megtartjatjuk a null-hipotézist.

A χ^2 próba gyakorlati megvalósításakor felosztjuk a számegyenest r részre ($x_1 < x_2 < ... < x_{r-1}$ osztópontokkal) úgy, hogy mindegyik részbe essen néhány (legalább 3) mintaelem. Ekkor az i-edik intervallum valószínűsége

$$p_i = P(X \in [x_{i-1}, x_i)) = F(x_i) - F(x_{i-1}) =$$

$$= 1 - \left(\frac{x_i}{\beta}\right)^{-\alpha} - \left(1 - \left(\frac{x_{i-1}}{\beta}\right)^{-\alpha}\right) = \frac{x_{i-1}^{-\alpha} - x_i^{-\alpha}}{\beta^{-\alpha}}$$

Ezt felhasználva a χ^2 statisztika kiszámolható, és összevethető az $(r-1)$-ed fokú χ^2 eloszlás megfelelő kvantilisével. Itt most ismét a folytonos eloszlásfüggvényt használtuk, de ez a χ^2 próbánál nem okoz gondot, mert az osztópontok megfelelő megválasztása esetén ezek a p_i értékek jól közelítik a valódiakat, és a próba szempontjából csak ezek a lényegesek.

A gráfok fokszámeloszlásának vizsgálatánál a kérdés néha fordított. Vagyis nem adott bizonyossági szint mellett akarunk dönteni a null-hipotézis elfogadásáról vagy elutasításáról, hanem éppen azt a legmasabb szignifikancia szintet keressük, amely mellett még elfoghatjuk H_0-t, vagyis hogy a vizsgált jelenség Pareto eloszlású.

3.2. A Küszöbkeresés algoritmus

Bizonyos gráfoknál csak az eloszlás lecsengő vége követi a Pareto eloszlást, az origóhoz közelebbi fele pedig valamilyen más eloszlásból származik (pl. Exponenciális, Gamma, Béta, Normális, stb.). Ilyen gráfokra később több példát is mutatunk. Ezeknél kíváncsiak vagyunk, hogy honnantól kezdve tekinthető a minta Pareto eloszlásúnak. Erre a következő módszert javasoljuk. Rendezzük növekvő sorba a mintaelemeket, és válasszuk ki a legnagyobb 5%-ot, 10%-ot, stb., a végén a teljes mintát. A tartományok tehát nem diszjunktak, hanem egymásba skatulyázottak. 5%-os lépésköz helyett természetesen választhatunk tetszőleges értéket, sőt a lépésközöknek nem is kell azonosaknak lenniük. Most legyenek a lépésközök azonosak és jelöljük m-el a felosztás finomságát, vagyis a tartományok számát. Végezzünk mindegyik tartományban becsléses χ^2 próbát, azaz az adott tartomány alapján először becslést adunk az (α, β) paraméterekre, majd illeszkedésvizsgálatot végzünk az említett módon (β értékére megközelítőleg x_i várható, ha az x_i-nél csonkítottunk). Amennyiben az eloszlás lecsengő vége Pareto eloszlású, úgy az 1. állítás miatt a legkisebb

tartománytól egy bizonyos tartományig, mondjuk a k-ig ($1 \leq k \leq m$) szintén Pareto eloszlást kell kapnunk, hiszen a „határ” felett minden csonkítás ismét Pareto eloszlást eredményez, méghozzá ugyanazzal az α paraméterrel. Tehát megkeressük azt a tartományt, amiben a minta még Pareto eloszlású (ε elsőfajú hibával), de a nála eggyel bővebben már nem. A végén ennek a tartománynak az alsó széle lesz az a határ, ahonnan kezdve Pareto eloszlásúnak tekinthető a minta, vagyis ahonnan a polinomiális lecsengés érvényesül. Ezt a módszert a Küszöbkeresés algoritmusnak hívjuk.

Algorithm 1 Küszöbkeresés

Require: X: mintaelemek sorozata
ε: megengedett elsőfajú hiba
m: osztópontok száma

$X \leftarrow$ Rendez(X)
$n \leftarrow$ Elemszám(X)
for $i = 1$ to m **do**
  $X_i \leftarrow \lfloor n \cdot \frac{m-i}{m} \rfloor$-edik értéknél nagyobbak X-ből
  $(\hat{\alpha}, \hat{\beta}) \leftarrow$ Becslés (X_i)
  if χ^2-Próba$(X_i, \hat{\alpha}, \hat{\beta}, \varepsilon) =$ **false then**
    return $\lfloor n \cdot \frac{m-i+1}{m} \rfloor$-edik érték X-ből
  end if
end for
return $\min X$

3.3. Egyenletes és Pareto eloszlás keveréke

Az algoritmus hatékonyságát eloszlások keverekékén teszteltük. Konstruáljuk a következő eloszlást: legyen az eloszlás 0 és 5 között egyenletes, 5 fölött pedig egy $\alpha = 2$, $\beta = 5$ paraméterű Pareto eloszlás. Mivel mindkét rész integrálja 1, ezért a sűrűségfüggvényeket még le kell osztani 2-vel, hogy valóban eloszlást kapjunk. Ebből az eloszlásból készítettünk egy nagyobb mintát, és futattuk rajta a Küszöbkeresés algoritmust úgy módosítva, hogy minden tartományt megvizsgáljon (és ne álljon le az első nem megfelelő tartománynál). Az eredményeket a 3. táblázatban foglaltuk össze. Amint az a táblázatból is kitűnik, az algoritmus igen pontosan találja meg

a küszüböt (a finomsága újabb osztópontokkal tovább növelhető). Az eredmények szerint a minta 4.9985 fölötti részére jól illeszthető a Pareto eloszlás.

Alsó határ	$\hat{\alpha}$	$\hat{\beta}$	OK
7.8963	1.9581	7.8943	1
6.5110	2.0040	6.5099	1
5.6073	1.9844	5.6066	1
4.9985	1.9748	4.9980	1
3.9886	1.5895	3.9882	0
3.0044	1.2451	3.0041	0
2.0248	0.9311	2.0246	0
1.0511	0.6341	1.0509	0
0.0001	0.0970	0.0001	0

3. táblázat. A Küszöbkeresés eredményei, 9 osztóponttal, $\varepsilon = 0.05$. Az OK oszlopban 1 szerepel, ha az alsó határ feletti rész Pareto eloszlásúnak tekinthető.

Valódi adatoknál sajnos nem mindig ilyen jó a helyzet. Ezeknél már sokkal jelentősebb az algoritmus pontatlansága, mert viszonylag nagy eltérések is lehetnek egy-egy helyen az elvárt és a megfigyelt értékek között. Ekkor enyhe módosításokat lehet eszközölni, pédául hogy megengedünk egy szeletnyi hibát. Vagyis ha az egyik tartományban elutasítjuk a Pareto-eloszlást, akkor még nem állunk le, csak ha a következő szeletnél is elutasítást kapunk. Ez azt jelenti, hogy egy 0-t még átugrunk az 1-esek sorozata között.

Az algoritmust úgy is módosíthatjuk hogy balról jobbra keressen, vagyis kiindulva a teljes tartományból, az alsó határt folyamatosan emelve az első olyan helyen álljon meg, ahol már elfogadható a Pareto-eloszlás. Ekkor tulajdonképpen az első 1-es jelzésű tartomány alsó határát adja vissza az algoritmus. Mi, hacsak mást nem említünk, ez utóbbit használjuk a következő fejezetekben. Azonban megemlítjük, hogy nem okozott lényegi különbséget az se, ha az előző, 1 hibát megengedő algoritmussal végeztük a számításokat.

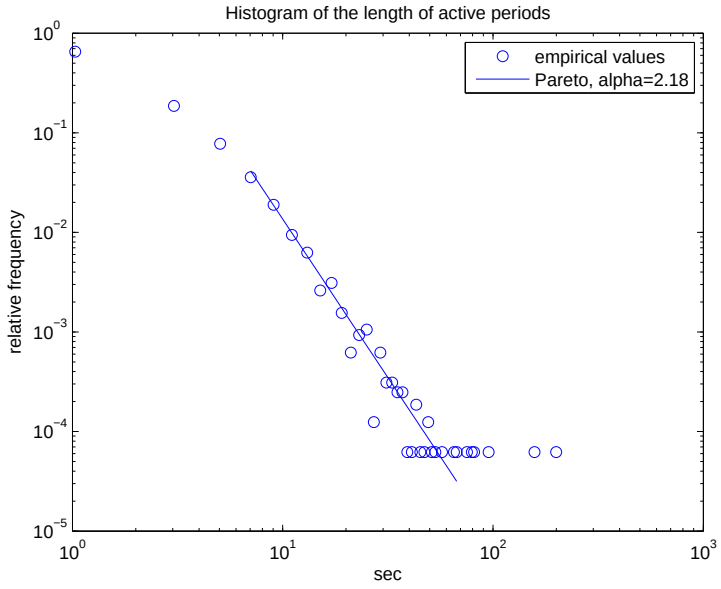
3.4. VoIP hívások vizsgálata

A bevezetőben említett és a következő fejetben részletesen bemutatott skálafüggetlen hálózatokon kívül sok más területen, például telekommunikációs jelenségeknél is találkozhatunk a Pareto eloszlással [2]. Napjainkban rohamosan terjed a VoIP, vagyis az IP hálózatokon keresztül lebonyolított telefonhívás. Egy VoIP kommunikáció során periódusok figyelhetőek meg [12]. Az egyik az úgynevezett *inaktív* periódus, amikor nincs nagy forgalom, és ami van, az is közel konstans mértékű. A másik az *aktív* periódus, amikor a két fél közötti forgalom megnő és dinamikusan változik, ingadozik. A kommunikáció során ilyen aktív és inaktív periódusok követik egymást. Kiderült, hogy ezen periódusok hossza is leírható Pareto eloszlással.

A Küszöbkeresés algortimussal megvizsgáltuk sok VoIP hívás összesen 16116 db aktív periódusának hosszát másodpercekben mérve. Az eredmények részlete a 4. táblázatban található. Itt látszik, hogy csak 6.83 fölött tekinthető a minta Pareto eloszlásúnak. Erre a felső régióra $\alpha = 2.1845$. A minta hisztogramja és a 2.18 paraméterű Pareto eloszlás elméleti sűrűségfüggvénye látható a 3. ábrán. Itt megfigyelhető a felső tartományokban a kiváló illeszkedés, továbbá az is, hogy a Pareto eloszlás sűrűségfüggvénye log-log skálázású ábrán egy egyenes vonal.

Alsó határ	$\hat{\alpha}$	$\hat{\beta}$	OK
9.4403	2.2614	9.4326	1
6.8394	2.1845	6.8365	1
5.4595	2.0336	5.4578	0
4.6208	1.9410	4.6197	0
3.9601	1.8174	3.9593	0

4. táblázat. A küszöbkeresés eredményeinek részlete aktív VoIP periódusok hosszain ($m = 30$, $\varepsilon = 0.01$).



3. ábra. VoIP hívások aktív periódusainak hisztogramja log-log skálázású ábrán, és az $\alpha = 2.18$ paraméterű Pareto eloszlás.

4. Gráffejlődési modellek vizsgálata

4.1. Skálafüggetlen hálózatok

Az egyik legfontosabb jelenség, ahol a Pareto avagy hatványfüggvény-eloszlás előfordul, a skálafüggetlen gráfok fokszámeloszlása. Az ilyen gráfokban annak a valószínűsége, hogy egy véletlenül kiválasztott csúcs fokszáma éppen egy k szám, k növekedtével polinomiálisan cseng le: $p(k) \sim k^{-\gamma}$. Erre utal a *skálafüggetlen* elnevezés, ugyanis a fokszámok hisztogramját log-log skálázású grafikonon ábrázolva egy egyenes vonalat láthatunk – függetlenül a logaritmus alapjától. Irányított gráfokra a skálafüggetlen tulajdonság a kimenő és a befutó élekre külön-külön is teljesül, általában különböző exponensekkel.

A témakörben az utóbbi években sok jelentős eredmény született, az alapokról jó összefoglaló [13]. Kiderült, hogy számos, a valódi világban fellelhető hálózat ilyen tulajdonságú. Néhány meglepő ezek közül:

- *WWW*: a gráf csúcsainak a weboldalak felelnek meg, az élek az oldalakon levő hiperlinkek
- *Hollywood*: a csomópontok a színészek, két színész közt akkor megy él, ha játszottak közös filmben
- *Tudományos kutatások*: a csomópontok a kutatók, élek a közös cikket jegyző kutatók között futnak
- *Szexuális kapcsolatok*: a csomópontok az emberek, közöttük akkor fut él, ha kapcsolatban voltak egymással

Az Internet is egy ilyen gráf, melyben számítógépek, routerek és egyéb eszközök vannak összekapcsolva. Az Internet topológiája két különböző szinten is vizsgálható. A *routerek* szintjén a csomópontok a routereknek illetve egyéb kommunikációs eszközöknek felelnek meg, melyek között a fizikai kapcsolat jelenti éleket. Az AS (*Autonomous System*) avagy *interdomain* szinten egy csomópont az egy szervezethez, domainhez tartozó routereket és számítógépeket takarja, ezek a hálózatok vannak összekötve egymással. Két csomópont akkor fut él, ha legalább egy vezeték megy köztük, vagyis van legalább egy-egy router mindkét domainben, amik össze vannak kötve egymással.

Mindkét szint fokszámeloszlásáról az elsők között derült ki, hogy hatványfüggvény alakú [15]. Mivel az AS szint határozza meg az Internet globális struktúráját, ezért célszerű ennek topológiáját felmérni illetve a kialakulásának folyamatát minél jobban modellezni. A precízebb modell segíthet a forgalom-elemzésben, a vírusok vagy támadások megállításában, stb.

Mint az a fenti felsorolásból is kitűnik, nem csak műszaki, hanem társadalmi-szociológia jelenségek között is sokhelyen előfordulnak skálafüggetlen hálózatok. Az emberek ismerősi kapcsolata is egy hálózat. Ebben a gráfban az emberek jelentik a csomópontokat és az egymást ismerő emberek vannak összekötve. Több szociológiai felmérés is igazolta, hogy ez a gráf, sőt az egy szervezeti vagy földrajzi egységre szűkett részgráfja is skálafüggetlen. Ennek fényében részletesen megvizsgáljuk egy magyar közösségi-site, az ismert `iWiW.hu` hálózatának egy részletét is, melyről azt tételezzük fel, hogy egy jó lenyomata az emberek tényleges ismerettségi hálózatának.

Arra kérdésre, hogy miként jönnek létre a skálafüggetlen hálózatok, hogyan jelenik meg a Pareto eloszlás, több modellt is javasoltak. Ezek közül néhány jelentősebbet mutatok be röviden.

Barabási-Albert (BA) modell

Ez az alap modell, melyet 1999-ben javasolt két amerikában élő magyar kutató, Barabási Albert-László és Albert Réka [1]. Két fontos eleme van. Az első, hogy a kezdeti m_0 darab ponthoz az idő múlásával újabb pontok adódnak hozzá, mondjuk időegységentként egy. A második, hogy minden új csúcs adott számú új élet alakít ki a korábban már a gráfban lévő csúcsokkal, és az hogy melyik korábbi ponthoz kapcsolódnak szívesebben, vagyis nagyobb valószínűséggel, az a pont fokszámával egyenesen arányos. Vagyis annak a valószínűsége, hogy egy új csúcs az i. régi csúcshoz kapcsolódik $p(i) = d(i)/\sum_j d(j)$ (ezt *lineáris preferenciának* is nevezik). Elméletileg és szimulációkkal is igazolták, hogy ez a fejlődési modell valóban skálafüggetlen hálózatot eredményez. A leírt folyamat következménye, hogy akinek sok kapcsolata van, annak az idők folyamán még több lesz („rich-get-richer" jelenség). Az alább ismertetett modellek ennek az általánosításai, bővítései.

Kibővített BA modell

A BA modell kibővítésében a gráf előző állapotához képest minden lépésben három lehetőség közül valamelyik történik [14]:

1. p valószínűséggel egy új él alakul ki a gráfban a már jelenlévő pontok között. Az él egyik vége véletlenszerűen választódik ki egyenletes eloszlás szerint, a másik vége a BA modellnél említett $p(i)$ valószínűségek szerinti lineáris preferenciával választódik ki.

2. q valószínűséggel egy él megszűnik, és helyette az él egyik végpontjától a szokásos fokszámok szerinti lineárispreferenciával egy másik csúcs fele jön létre egy új él.

3. $1 - p - q$ valószínűséggel egy új pont lép be, és kapcsolódik a szokásos preferenciás választással néhány régi csúcshoz, mint az egyszerű BA modellben.

Generalized Linear Preference (GLP) modell

A GLP modellben egyrészt a lineáris preferencia általánosításaként minden új él kialakulásakor az i. ponthoz való kapcsolódás valószínűsége $p(i) = (d(i) - \beta)/\sum_j (d(j) - \beta)$, ahol $\beta \in (-\infty, 1)$ állítható paraméter. Ez az érték a kis pontok „hátrányát" csökkenti némileg a nagyokhoz képest, de valójában nem okoz számottevő különbséget, mert előbb utóbb úgyis lesznek olyan pontok, akik jóval több élre tesznek szert, és így jóval preferáltabbakká válnak. Másrészt, mint a kibővített BA modellben, így itt is minden körben p valószínűséggel két korábbi pont között alakul ki él, új pont csatlakozása nélkül. A GLP modellben azonban nem szűnnek meg élek, ezért tulajdonképpen nagyon hasonló a $q = 0$ paraméterű kibővetett BA modellhez.

Interactive Growth (IG) modell

Ezt a modellt kifejezetten az AS gráf minél jobb reproduálására alkották [16]. Egy új csomópont itt is a BA modellben bemutatatott lineáris preferenciával választja meg, hogy mely régi csomópontokhoz kapcsolódjon. Ezután viszont azoknak a csomópontoknak, amelyekhez ez az új csomópont elsőként csatlakozott, szintén van lehetősége más csomópontokhoz kapcsolódni ugyanúgy lin. pref. szerint. Ennek magyarázata az AS gráf esetén az, hogy ezek a domainaek csillapítani akarják az új

csomópont által gerjesztett forgalmat. A szerzők a következő paramétereket javasolják: egy új pont 0.6 valószínűséggel kapcsolódhat 2 db illetve 0.4 valószínűséggel 1 db élhez. Előbbi esetben a két régi csomópont közül az egyik még kaphat egy új élet, utóbbi esetben pedig az egy darab régi csomópont kaphat még két élet.

4.2. A modellek statisztikai tulajdonságai

Egy korábbi tanszéki fejlesztésnek köszönhetően lehetőség nyílt a BA, GLP és IG modellek szerint gráfokat generálni. Mindhárom modellből 50–50 db, 7368 pontú gráfot generáltunk. Mint látni fogjuk, ez az általunk mért AS gráf csúcsainak száma. Ezután mindegyik gráfra megkerestük a küszöböket, a kisebb értékek felől elvégezve a Küszöbkeresést az első illeszkedő tartományig. Ezekre a küszöbértékekre mostantól TH-ként (treshold) is hivatkozunk. Az értékeknek a minimumát, maximumát és átlagát az egyes típusokon belül a 5. táblázat foglalja össze. Az adatokból leolvasható, hogy mindhárom esetben általában 4–9 körül kezdődik az az érték, ami fölött érvényesül a skálafüggetlen tulajdonság, vagyis a Pareto eloszlás. A BA modellben ez viszonylag szűk tartományon belül ingadozik, az IG modellnél viszont nagyon ritka esetekben akár 30 fölötti is lehet.

Modell	TH_{min}	TH_{max}	TH_{avg}
BA	4	9	4.94
GLP	6	14	8.9
IG	3	34	5.14

5. táblázat. Küszöbértékek minimuma, maximuma és átlaga 50-50 futás eredményei alapján ($m = 50$, $\varepsilon = 0.01$).

A legmagasabb TH_{avg} érték a GLP modellnél adódott, viszont a gráfok nagyrészére 15-nél kisebb értéket kaptunk. Így ezt vettük egy olyan határnak, amely fölött a generált gráfok legtöbbje már hatványfüggvény eloszlásúnak tekinthető. Az e fölötti tartományra minden gráfon egy α becslés végezhető, mely jó összehasonlítási alapot ad a különböző modellek között. A becsült α-k átlagai és a becslések szórásai a 6. táblázatban láthatóak. Barabásiék mean-field módszerrel elméleti úton belátták, hogy modelljukben az α értéke 2 körül kell legyen. Ez nagyrészt összhangban áll a mi empirikus eredményeinkkel, mert egy kicsit kisebb, 1.92 körüli értéket kaptunk

átlagként. A GLP és IG modellekre tudomásunk szerint nem áll rendelkezésre ilyen elméleti eredmény.

Modell	α_{avg}	α_σ
BA	1.9292	0.1530
GLP	1.8448	0.0372
IG	1.2714	0.0451

6. táblázat. Az $x \geq 15$ fokszámokra becsült α értékek átlaga és szórása, 50-50 futás eredményei alapján ($m = 50$, $\varepsilon = 0.01$).

4.3. Az AS gráf

Mint említettük, az AS gráf feltérképezésére, kialakulásának megértésére manapság nagy igény mutatkozik. Erre a hálózatra a kibővített BA modell legfeljebb $q \cong 0$ paraméterrel lenne reális, mert az a legritkább esetben fordul elő, hogy egy régi kapcsolatot megszűntetnek, általában inkább újabb, gyorsabb vonalakat alakítanak ki a már meglévők mellé. Ekkor viszont a GLP modellt kapjuk vissza $\beta = 0$ paraméterrel.

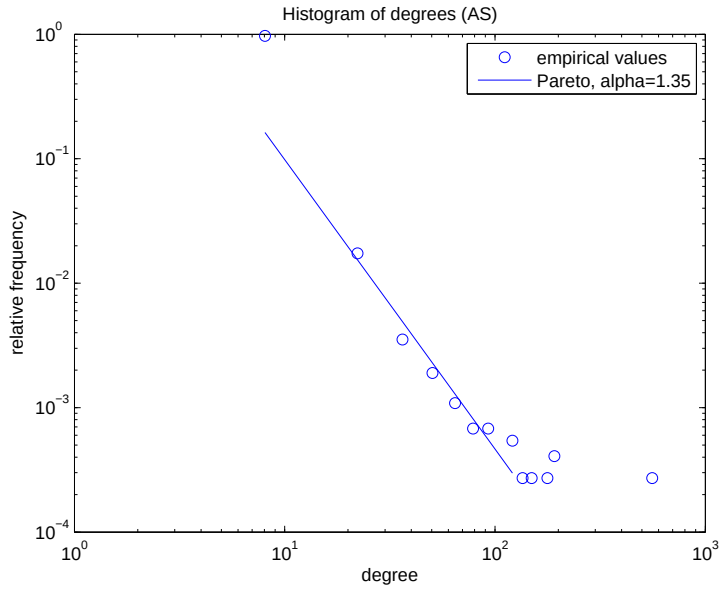
Az előző fejezetekben kimunkált statisztikai módszerekkel mi is vizsgálatokat végeztünk a 2007. októberi AS hálózaton [17], mely mérésünk időpontjában egy 7368 csúcsú gráf volt. A Küszöbkeresés algoritmus kimenetének részlete a 7. táblázatban látható. Látható, hogy az AS gráfot nagyjából az $x \geq 7$ értékre tekinthetjük Pareto eloszlásúnak, e fölött az $\alpha = 1.346$, ami a 4. ábrán is látható. A bevezetett közös küszöb feletti, vagyis az $x \geq 15$ értékekre az $\alpha = 1.3215$ paraméterű görbe illeszkedik a legjobban. Ez összhangban van a korábbi mérésekkel [15].

Az α paraméter alapján az AS gráf a legjobb egyezést az IG modellel mutatja, hiszen az így generált gráfokra ennek átlaga 1.2714, ez az említett 1.3215-ös értékhez képest 4%-os eltérést jelent. Ez nem meglepő, hiszen az IG modellt kifejezetten az AS szint kiépülésének feltételezett mechanizmusai alapján terveztek.

Zhou és Mondragón, az IG modell megalkotói [16], azt is kimutatták, hogy az AS és egy IG gráf az úgynevezett „rich-club" jelenségben is kiválóan megegyeznek. Ez azt jelenti, hogy ha a gráfból kiválasztjuk a legnagyobb fokú pontokat, akkor

Alsó határ	$\hat{\alpha}$	$\hat{\beta}$	OK
15.0000	1.3215	14.9525	1
12.0000	1.3502	11.9724	1
9.0000	1.3457	8.9855	1
7.0000	1.3466	6.9917	1
6.0000	1.3633	5.9943	0
5.0000	1.4158	4.9965	0

7. táblázat. A küszöbkeresés eredményeinek részlete az AS gráfon ($m = 100$, $\varepsilon = 0.01$).



4. ábra. Az AS gráf fokszámainak hisztogramja log-log skálázású ábrán, és az $\alpha = 1.35$ paraméterű Pareto eloszlás.

közöttük az összes lehetséges él közül mennyi van meg tényleg a gráfban. A valódi AS gráfnál ez igen magas, vagyis a gazdag pontok egy „klubbot" alkotnak, szinte mindegyik össze van kötve mindeggyikkel. Az IG modellben a kiválasztott pontok számának függvénye nagyon hasonló a valódi AS gráf ugyanilyen függvényéhez.

4.4. Az iWiW gráf

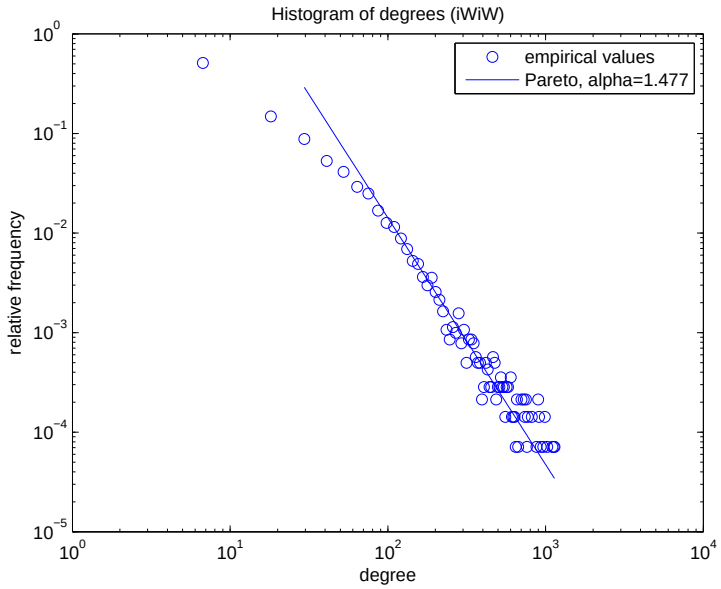
Az `iWiW.hu` mára Magyarország legnagyobb közösségi site-ja lett. A rendszer használói bejelölik ismerőseiket, így alakítva hálózatot. Az AS gráfhoz hasonlóan az iWiW gráf egy 14060 pontú részgráfjának illeszkedését is megvizsgáltuk. A részgráfot szélességi bejárással vágták ki, vagyis egy emberből kiindulva vették a szomszédait, majd az ő szomszédaikat, stb. 2-3 szintig (pontosan mi sem tudjuk, milyen mélységig, de ez most nem is lényeges).

Az így kapott iWiW részgráfnál több érdekes dolgot tapasztaltunk. Először is az illeszkedés alsó határa meglepően magas, körülbelül 60. Ez a 5. ábrán is látszik, ami a gráf fokszámainak hisztogramját mutatja log-log slálázású diagrammon. Az eloszlás lecsengő vége kiválóan illeszkedik a hatványfüggvény eloszlásra (egyenes vonal), de kb. 60 körül egy törést láthatunk az ábrán, és a küszöbnél kisebb értékekre messze elmarad a valódi gyakoriság az elvárttól. Meglepő módon a törés alatti rész is egy egyenes a log-log ábrán, ami azt jelenti, hogy ez is hatványfüggvény alakú, csak más paraméterekkel. Az α például biztosan kisebb itt, mert a vonal is kevésbé lejt.

A küszöbkeresést jóval kevesebb osztóponttal futtatuk, mivel a gráf elég nagy. Ezért a kapott eredmény feltehetőleg pontatlanabb a korábbiaknál. A 8. táblázatban található az eredmények egy részlete. Látható, hogy 61-et kaptunk alsó határnak, és e fölött az $\alpha = 1.4776$.

Alsó határ	$\hat{\alpha}$	$\hat{\beta}$	OK
134.0000	1.7202	133.8897	1
84.0000	1.5915	83.9627	1
61.0000	1.4776	60.9807	1
46.0000	1.3358	45.9880	0
36.0000	1.2153	35.9917	0

8. táblázat. A küszöbkeresés eredményeinek részlete az iWiW gráfon ($m = 20$, $\varepsilon = 0.01$).

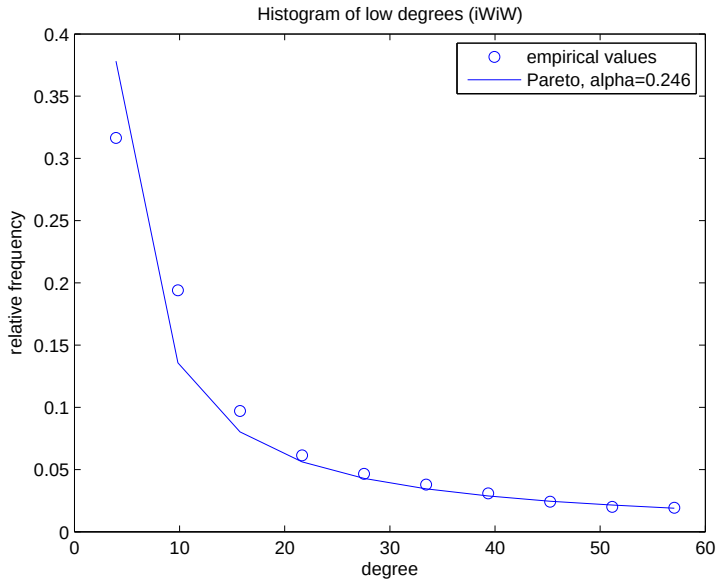


5. ábra. Az iWiW gráf fokszámainak hisztogramja log-log skálázású ábrán, és az $\alpha = 1.477$ paraméterű Pareto eloszlás.

Ha csak az kis értékekre akarunk Pareto eloszlást illeszteni, akkor nem használhatjuk az eddigi CML becslést, ugyanis így a valódi α-nál jóval kisebb értéket kapnánk, a nagy fokszámok hiánya miatt. Viszont az LSF becslés használható. Ez ugyanis a hisztogramra illeszti a görbét, és ehhez elég a kis minták ismerete is. A 6. ábrán látható, hogy az $\alpha = 0.246$ paraméterű eloszlás megfelelő mértékben illeszkedik a pontokra. Ez meglepően kicsi a felső régiók 1.477-es paraméteréhez képest, és azt mutatja, hogy az 50-60-nál kisebb fokszámú pontok másképp viselkednek, mint a nagyobb fokszámúak.

Azt, hogy a kis fokszámú pontokból kevesebb van a vártnál, nem magyarázza meg a részgráf kivágásának módszere. Ugyanis a szélességi bejárás utolsó szintjén bevett pontoknak még nincs is minden éle a részgráfba felvéve, amikor leáll a bejárás, így elvileg a kis fokszámú pontoknak sokkal gyakoribbnak kellene lenniük a részgráfban, mint a teljes gráfban. Kérdés tehát, hogy mi okozhatja a megtörési a jelenséget.

Ez talán azért lehet, mert amikor egy ember csatlakozik az iWiW-hez, akkor egy kezdeti fázisban intenzíven elkezd kapcsolatokat építeni. Ebben a kezdeti fázisban



6. ábra. Az iWiW gráf 60-nál kisebb fokszámainak hisztogramja egyenletes skálázású ábrán, és a rá illesztett $\alpha = 0.246$ paraméterű Pareto eloszlás.

felkutatja a barátait, így a csatlakozás után kevés idővel már sok kapcsolata van. A barátok utáni kutatás általában úgy történik, hogy az ember szétnéz az eddigi ismerőseinek az ismerősei közt, és sokszor közülük sokakat ismer, hiszen egy társaságból valók. Ekkor hirtelen sok új kapcsolatra is szert tehet, vagyis „lemásolhatja" egy nagy fokszámú pont éleit, így ő is viszonylag könnyen nagy fokszámúvá válhat. Ebből kifolyólag az elvártnál kevesebb lesz a kis fokszámú pont. Ezen kezdeti fázis után azonban elfogy a lendülete és „lassítani" kezd, így a nagyobb régiókban lassab lesz a bővülés. Ez természetesen csak egy heurisztikus felvetés, további szimulációk szükségesek a fent említett hatások helyességének eldöntésére.

5. Az elkészült programok bemutatása

A dolgozathoz készített `.m` fájlok Matlab-ban futtathatóak, és jellegüknél fogva két csoportra oszthatóak. Az egyik csoport az algoritmusokat valósítja meg. Ezek mindig paraméterként kapják meg a bemenetet, és a kimenetet vagy kiírják, vagy visszatérési értékként adják tovább. A másik csoport a dolgozatban leírt kísérleteket automatizálja, ezeknek nincs semmilyen paraméterük vagy visszatérési értékük, így egyszerűen futtathatóak. Amennyiben ezekben a paraméterek állítására lenne szükség, azt a fájlok szerkesztésével lehet megoldani. Mindegyik fájlhoz egy rövid fejkomment is tartozik, ebből kiderülnek az esetleges paraméterek és visszatérési értékek.

Algoritmusok

`chi2t.m`	χ^2-próba egy megadott Pareto eloszlásra
`cmle.m`	korrigált maximum likelihood paraméterbecslés
`lsf.m`	paraméterbecslés lineáris regresszióval
`mle.m`	maximum likelihood paraméterbecslés
`moe.m`	paraméterbecslés momentum-módszerrel
`tme.m`	paraméterbecslés csonkított momentum módszerrel
`treshold_search.m`	küszöbkeresés (alap verzió)
`treshold_search1.m`	küszöbkeresés (1 hibát megengedő verzió)
`treshold_searchl.m`	küszöbkeresés balról kezdve
`treshold_searchp.m`	küszöbkeresés (a teljes táblázat kiírásával)

Kísérletek, mérések

`caida.m`	kísérletek az AS gráfon (4.3 fejezet)
`graph_stat.m`	50-50 generált gráf statisztikái (4.2 fejezet)
`mixgen.m`	egyenletes és Pareto eloszlás keveréke (3.3 fejezet)
`testalpha.m`	α becslések összehasonlítása (2.3 fejezet)
`testbeta.m`	β becslések összehasonlítása (2.3 fejezet)
`tm_stat.m`	a csonkított mom. módszer adaptációja (2.2.2 fejezet)
`voip.m`	kísérletek VoIP adatokon (3.4 fejezet)
`wiw.m`	kísérletek az iWiW gráfon (4.4 fejezet)

6. Összefoglalás, konklúzió

Ebben a dolgozatban részletesen megvizsgáltam a skálafüggetlen hálózatok fokszámeloszlását, paramétereit felderítő módszereket. Először bemutattam az elterjedt paraméterbecslő módszereket a Pareto eloszlásra (lineáris regresszió, momentum módszer, maximum likelihood módszer). Ezek mellett bemutatásra került egy kevésbé ismert becslési módszert, mely bizonyos csonkított momentumok eloszlás-karakterizáló tulajdonságán alapul. Mérési eredmények igazolják, hogy ezen becslési módszerek közül a Saksena és Johnson által [7]-ben levezetett CMLE vagyis korrigált maximum likelihood becslés a legoptimálisabb. Mindazonáltal más eloszlások esetén érdemes megfontolni a bemutatott TME módszer használatát is, különösen abban az esetben, ha a maximum likelihood egyenlet csak nehezen, vagy csak numerikusan lenne megoldható.

Az TME módszer alkalmazásakor kiderült, hogy a csonkítási pontokat magukban a mintaelemekben érdemes kijelölni, és a csonkítás után alkalmazott regressziókor a pontoknak csak az alsó 50-70%-át érdemes figyelembe venni. A TME módszer a bemutatotthoz hasonló módon alkalmazható minden olyan eloszlásra, melyek karakterizáló egyenletei lineáris regresszióra visszavezethetők. További kutatás vagy általánosított regresszió szükséges az olyan eloszlások esetére, ahol ez nem lehetséges.

Illeszkedésvizsgálatra bemutattam egy küszöbkereső algoritmust, amely megtalálja, hogy mely érték fölött tekinthető a vizsgált minta Pareto eloszlásúnak. Ez különböző módosításokkal is futtatható, mi a legengedékenyebb megoldást választottuk. Ennek segítségével megvizsgáltuk VoIP hívásokban az aktív peiódusok hosszát. Kiderült, hogy a magasabb értékekre jól illeszkedik az Pareto eloszlás $\alpha = 2.18$ paraméterrel.

Napjainban a Pareto eloszlás a skálafüggetlen gráfok vizsgálatakor kerül elő leggyakrabban. Ezért statisztikai módszereink legfőbb alkalmazási területe ezen gráfok vizsgálata. Ennek jegyében megvizsgáltam néhány gráffejlődési modellt (BA, GLP, IG), és összehasonlítottam őket valódi hálózatokkal, pontosabban az AS és az iWiW gráffal. Ezen vizsgálatok során kiderült, hogy az IG modell kiválóan illeszkedik az AS gráfra, mindkettőben az alsó küszöb 5-10-es nagyságrendű és e fölött az α paraméter értéke 1.3 körül van.

Az iWiW hálózatra viszont egyik fenti modell sem illeszkedik igazán, egy viszonylag magas fokszámnál lévő törés ugyanis két részre bontja az értéktartományt.

A törés alatt és fölött is hatványfüggvény alakú az eloszlás, de más-más exponensekkel. Feltételezésem szerint az iWiW gráf fejlődésekor más szempontok is szerepet játszanak, mint a többi hálózatnál.

Természetesen a skálafüggetlen hálózatokat sok más szempont alapján is lehet vizsgálni, ezek közül nagyon sokat már kidolgoztak a kutatók. Dolgozatomban a Pareto eloszlás illesztésére helyeztem a hangsúlyt, mely véleményem szerint igen érdekes feladat volt. Remélem, hogy ezen eszközök segítségével a későbbiekben a skálafüggetlen hálózatok még részletesebb megismerésére is mód nyílik.

A. Statisztikai emlékeztető

Egy X valószínűségi háttérváltón megfigyeléseket végzünk. Ekkor *statisztikai minta* alatt a független, azonos eloszlású $X_1, X_2, \ldots$ valószínűségi változók sorozatát értjük, ahol az X_i valószínűségi változük eloszlása megegyezik az X háttérváltozóéval. A megfigyelések egy konkrét *realizációját* $x_1, x_2, \ldots$ jelöli. A mintaelemek egy $T(X_1, \ldots, X_n)$ függvényét *statisztikának* nevezzük.

Alapstatisztikák

Mintaátlag

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Empirikus szórásnégyzet

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Korrigált empirikus szórásnégyzet

$$S^{*2} = \frac{n}{n-1} S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

K-adik empirikus momentum

$$M_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

Becsléselmélet

Most feltesszük, hogy a háttérváltozó valamilyen θ paraméterrel paraméterezhető eloszláscsaládból származik.

1. Definíció. *A **likelihood-függvény** a mintaelemek együttes sűrűségfüggvényét (diszkrét esetben súlyfüggvényét) jelenti.*

$$L_\theta(x) = \prod_{i=1}^{n} f_\theta(x_i)$$

2. Definíció. *A $T(X)$ statisztika* ***elégséges*** *a θ paraméterre ha az*

$$f_\theta(x|T(X)=t) = \begin{cases} \frac{L_\theta(x)}{f_\theta^T(t)}, & \text{ha } T(X)=t, \\ 0, & \text{különben} \end{cases}$$

feltételes sűrűségfüggvény nem függ θ-tól, ahol $f_\theta^T(t)$ jelöli a $T(X)$ statisztika sűrűségfüggvényét a t helyen, $L_\theta(x)$ pedig a minta likelihood-függvényét.

Ez definició a folytonos esetben érvényes. Diszkrét eloszlásra hasonló, csak sűrűségfüggvény helyett súlyfüggvénnyel.

3. Definíció. *A T statisztika* ***teljes****, ha*

$$\mathrm{E}(g(T)) = 0, \quad \forall\theta \Longrightarrow g = 0 \text{ majdnem mindenütt}$$

Vagyis a teljes statisztikának csak az azonosan 0 függvénye lesz 0 várható értékű.

Az elégségesség és teljesség fogalma is tulajdonképpen a T statisztikának a θ paraméterre vonatkozó információ-megőrzését fejezi ki. Amíg az elégséges statisztika a mintaelemekből minden θ-ra vonatkozó információt megőriz, addig a teljes statisztika semmilyen olyan információt nem őriz meg, ami nem a θ-ra vonatkozik. Az elégséges statisztikák között bevezethető egy részben rendezés, „információ-gazdaságossági" szempont alapján: T_1 a T_2-nek alárendelt statisztika, ha létezik ν függvény, hogy $T_1 = \nu(T_2)$.

4. Definíció. *A T statisztika* ***minimális elégséges****, ha alárendelt statisztikája bármely más elégséges statisztikának.*

2. Tétel. *Ha a T elégséges statisztika teljes, akkor minimális elégséges is.*

Most a θ paramétert, vagy annak valamilyen $\psi(\theta)$ függvényét szeretnénk becsülni az $X = (X_1, \ldots, X_n)$ független azonos eloszlású minta alapján konstruált $T(X)$ statisztika segítségével.

5. Definíció. *$T(X)$* ***torzítatlan*** *becslés $\psi(\theta)$-ra, ha*

$$\mathrm{E}(T(X)) = \psi(\theta), \quad \forall\theta$$

6. Definíció. *Legyen T_1 és T_2 torzítatlan becslés (θ-ra vagy valamilyen $\psi(\theta)$-ra). T_1* ***hatásosabb*** *T_2-nél, ha*

$$\mathrm{D}^2(T_1) \leq \mathrm{D}^2(T_2), \quad \forall\theta$$

és legalább egy θ_0 esetén határozott egyenlőtlenség teljesül.

7. Definíció. *Egy torzítatlan becslés* ***hatásos****, ha bármely más torzítatlan becslésnél hatásosabb.*

3. Tétel (Lehmann-Scheffé). *Ha T elégséges, teljes és torzítatlan, akkor hatásos.*

Ez a tétel a Rao-Blackwell-Kolmogorov-tétel következménye [8], utóbbi bizonyítása megtalálható [4]-ben.

Hivatkozások

[1] Barabási, A.-L., Albert, R., Emergence of Scaling in Random Networks, *Science* 286, 509-512 (1999).

[2] Newman, M. E. J., Power-laws, Pareto distributions and Zipf's law, *Contemporary Physics* 46, 323-351 (2005).

[3] Goldstein, M. L. et al., Problems with fitting to the Power-Law Distribution, *The European Physical Journal B* 41, 255-258 (2004).

[4] Bolla M., Krámli A., *Statisztikai következtetések elmélete*, Typotex, Budapest (2005).

[5] Weisstein, E. W., "Least Squares Fitting–Power Law", *MathWorld*, `http://mathworld.wolfram.com/LeastSquaresFittingPowerLaw.html`

[6] Malik, H. J., Estimation of the parameters of the Pareto distribution, *Metrika* 15, 126-132 (1970).

[7] Saksena, S. K., Johnson, A. M., Best unbiased estimators for the parameters of a two-parameter Pareto distribution, *Metrika* 31, 77-83 (1984).

[8] Wikipedia contributors, "Rao–Blackwell theorem," *Wikipedia, The Free Encyclopedia*, `http://en.wikipedia.org/Rao-Blackwell_theorem`

[9] Glänzel, W., A characterization theorem based on truncated moments and its application to some distribution families, *Mathematical Statistics and Probability Theory* B, 75-84 (1987).

[10] Hamedani, G. G., Characterizations of Cauchy, Normal and Uniform Distributions, *Studia Scientenarium Mathematicarum Hungarica* 28, 243-247 (1993).

[11] Hamedani, G. G., Glänzel, W., Characterization of Univariate Continous Distributions, *Studia Scientenarium Mathematica Hungarica* 37, 83-118 (2001).

[12] Dang T. D., Sonkoly B., Molnár S., Fractal Analysis and Modeling of VoIP Traffic, *11th International Telecommunications Network Strategy and Planning Symposium* (2004).

[13] Albert, R., Barabási, A-L., Statistical Mechanics of Complex Networks, *Reviews of Modern Physics* 47, 74 (2002).

[14] Albert, R., Barabási, A-L., Topology of Evolving Networks: Local Events and Universality, *Physical Review Letters* 85, 24 (2000).

[15] Faloutsos, M., Faloutsos, P., Faloutsos, C., On Power-Law Relationships of the Internet Topology, *ACM SIG-COMM '99* (1999)

[16] Zhou, S., Mondragón, R. J., Towards modelling the Internet topology - the Interactive Growth model, *Proc. of the 18th International Teletraffic Congress* (2003).

[17] The CAIDA AS Relationships Dataset, 2007.10.02. `http://www.caida.org/data/active/as-relationships/`