Statisztikai módszerek a skálafüggetlen hálózatok vizsgálatára

Gyenge Ádám[1]

[1]Budapesti Műszaki és Gazdaságtudományi Egyetem
Villamosmérnöki és Informatikai Kar
Számítástudományi és Információelméleti Tanszék

2007. november 17.

Konzulens: Dr. Telcs András

Skálafüggetlen gráfok

- Nagyméretű hálózatok
- Fokszámeloszlásuk hatványfüggvény alakú

$$p(k) \sim k^{-\gamma}$$

- Rengeteg helyen előfordulnak. Pl.

Hálózat	csúcsok	élek
WWW	weboldalak	hiperlinkek
Hollywood	színészek	közös film
Tudományos kutatások	kutatók	közös cikk

- Cél: a fokszámeloszlás eddigieknél precízebb vizsgálata (ez sok tulajdonságot meghatároz)
- Ezáltal jobb modellezés ($\rightarrow$ pl. szimulációk: vírusok, támadások, pletykák,...)

Pareto eloszlás

Sűrűségfüggvénye:

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

Illesszünk Pareto eloszlást a fokszámokra!

Paraméterek (amiket keresünk)

- α: lecsengés gyorsasága (ez a lényegesebb)
- β: távolság az origótól

Paraméterbecslés - klasszikus módszerek

- Visszavezetés lineáris regresszióra:

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

  Hisztogramfüggvény loglog ábrán $\rightarrow$ erre egy egyenes illesztése
- Momentum módszer: a $H : (\alpha, \beta) \longrightarrow (m_1, m_2)$ leképezés invertálása
- Maximum likelihood módszer:

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n} \ln x_i - n \ln x_1^*} \qquad \hat{\beta} = x_1^* = \min\, x_i$$

- Ezek mind torzított becslések.
- A maximum likelihood torzítatlanná tehető $\rightarrow$ korrigált maximum likelihood becslés, CMLE:

$$\hat{\alpha} = \left(\frac{n-2}{n}\right) \cdot \hat{\alpha}_{MLE}$$

Csonkított momentumok módszere

Tétel (Glänzel)

Minden abszolút folytonos eloszlású X valószínűségi változóra megfelelő g, h és λ függvényekkel igaz, hogy:

$$E(g(X)|X \geq x) = E(h(X)|X \geq x)\lambda(x)$$

Következmény: a Pareto eloszlásra igaz hogy $1 = \frac{\alpha+1}{\alpha} \overbrace{E(X^{-1}|X \geq x)}^{v(t)} x$
A $v(t)$ függvény becslése egy t helyen:

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

Ezt vegyük fel a $\underline{t} = (t_1, \ldots, t_n)$ pontokban $\rightarrow \underline{v}$
Erre: $1 - \frac{\alpha+1}{\alpha} \underline{v} \cdot \underline{t} \approx 0 \rightarrow$ lineáris regresszióval megtalálható az együttható

A becslési módszerek összehasonlítása

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

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

Látható, hogy a CMLE a legpontosabb.
A β becslésekre is hasonló eredményeket tapasztaltunk (a CMLE volt a legjobb).

Illeszkedésvizsgálat

Előfordulhat, hogy csak az eloszlás lecsengő vége követi a hatványfüggvényt (példák később)

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.

Adott mintára egy t pontbeli csonkítás: a t-nél kisebb elemek elhagyása

Az állítás miatt az α változatlan a felsőbb tartományokra.

Küszöbkeresés algoritmus: meghatározza, hogy milyen küszöb felett illeszthető Pareto eloszlás a mintára (adott bizonyossággal)

Alkalmazás: VoIP forgalom

VoIP kommunikáció: *inaktív* és *aktív* periódusok váltakoznak
Aktív periódusok hosszaira vonatkozó mérés részletei:

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

Eredmény: 6.8 sec fölött $\alpha = 2.18$ paraméterű Pareto

Gráffejlődési modellek

A skálafüggetlen gráfok kialakulására rengeteg modellt javasoltak.
Néhány ezek közül:

1. BA modell
   - folyamatosan növekvő hálózat
   - az i. régi csúcshoz való kapcsolódás valószínűsége $p(i) = d(i)/\sum_j d(j)$
2. GLP modell
   - $1 - p$ valószínűséggel egy új pont lép be
   - p valószínűséggel két régi pont között épül fel új kapcsolat
   - az i. csúcshoz való kapcsolódás valószínűsége $p(i) = (d(i) - \beta)/\sum_j (d(j) - \beta)$, $\beta \in (-\infty, 1)$
3. IG modell
   - egy új pont 1 vagy 2 régi ponthoz kapcsolódhat lineáris preferencia szerint
   - ez az 1 vagy 2 pont új kapcsolatokat alakíthat ki (az új pont által generált forgalom elvezetése céljából)

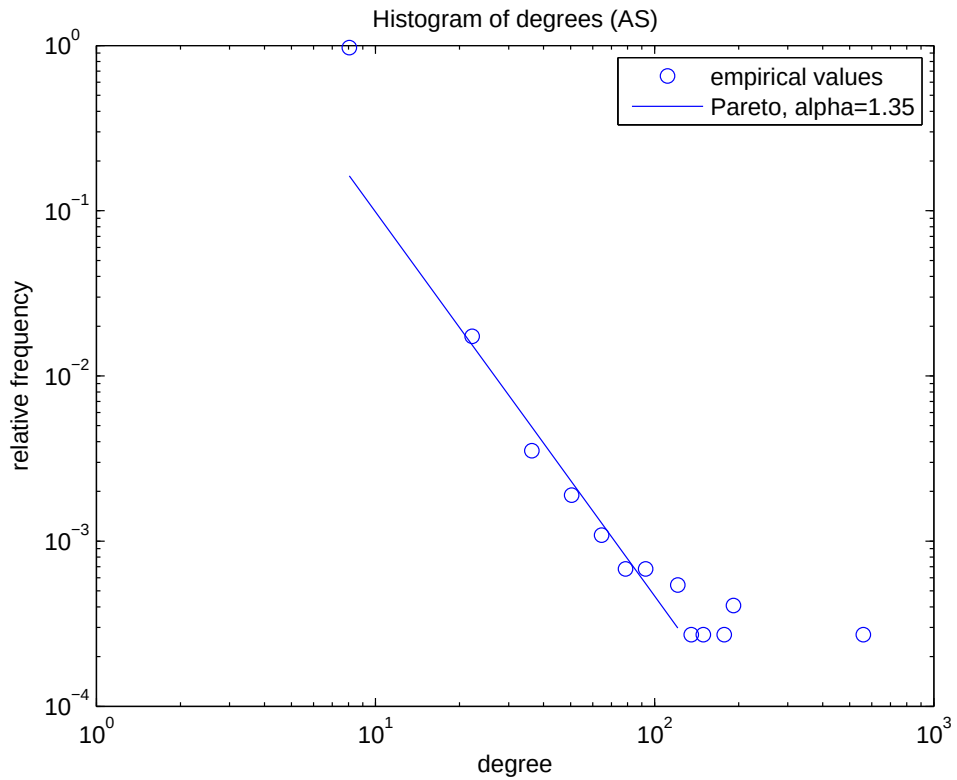
A modellek tulajdonságai

- Mindhárom modell szerint 50-50 nagyméretű ($n = 7368$) generált gráf.
- Átlagos küszöbértékek: BA - 4.94, GLP - 8.9, IG - 5.14
- A 15 feletti fokszámokra szinte mindegyik gráfban teljesült a Pareto eloszlás
- A 15 feletti tartományok α értékei:

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

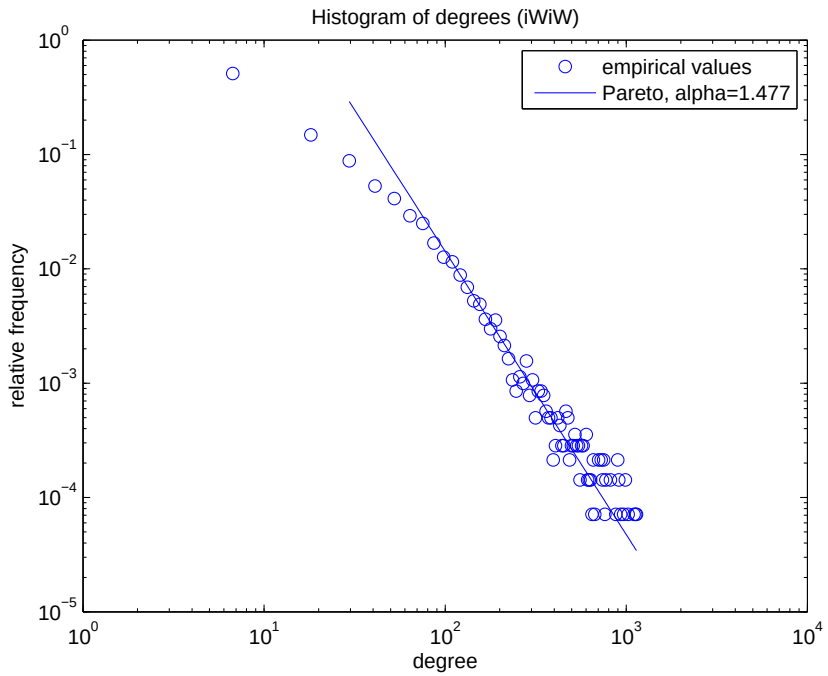
AS gráf

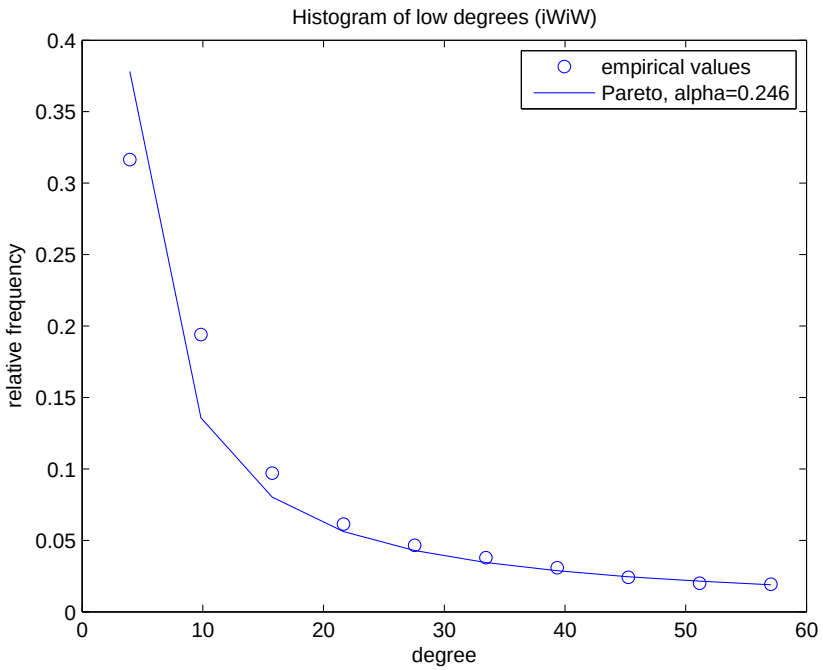
Internet: hálózatok összekapcsolva
AS gráf: csomópont = hálózat, él = fizikai kapcsolat a két hálózat között
2007. októberben 7368 pontú gráf
Küszöb: 7 körül, e felett $\alpha = 1.346 \rightarrow$ IG kiválóan modellezi
Az AS gráf fokszámainak hisztogramja:



iWiW gráf

iWiW: közösségi site
Ennek egy 14060 pontú részgráfját vizsgáltuk
Küszöb: 60 körül (!), e felett $\alpha = 1.477$
Ez meglepően magas, egyik bemutatott modell sem generál hasonlót
Másik meglepő eredmény: a küszöb alatt is hatványfüggvény alakú az eloszlás, de $\alpha = 0.246$





Összefoglalás

- Láthattuk hogy a skálafüggetlen hálózatok és a Pareto eloszlás rengeteg helyen előfordulnak
- Paraméterbecslésre a CMLE módszer a legjobb
- Küszöbkeresés algoritmussal megtalálható a Pareto eloszlás alsó határa
- VoIP kommunikáció aktív periódusaira a küszöb kb. 6 sec, $\alpha = 2.18$
- Az AS gráfban a küszöb 7, $\alpha = 1.34$ és az IG modell erre jól illeszkedik
- Az iWiW gráfban a küszöb 60, $\alpha = 1.47$
- De ez utóbbinál a küszöb alatt is hatványfüggvény alakú az eloszlás, $\alpha = 0.245$

Köszönetnyilvánítás

Szeretnék köszönetet mondani:

- Dr. Telcs Andrásnak
- Somogyi Róbertnek
- Molnár Sándornak
- Balogh Miklósnak

Köszönöm a figyelmet!

Kérdések?