Rast informačných sietí

Matúš Medo, Giulio Cimini, Stanislao Gualdi

University of Fribourg, Switzerland

Trojkráľová konferencia 2013
Banská Bystrica

4. január 2013

Žijeme v dobe informačnej

Žijeme v dobe informačnej

Informácia je vytváraná, šíri sa, je zabudnutá

Žijeme v dobe informačnej

Informácia je vytváraná, šíri sa, je zabudnutá

this talk

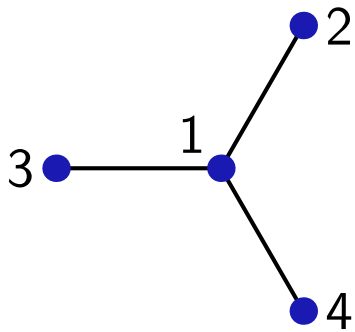
Preferential attachment (PA)

- Klasický model komplexnej siete
  - Yule (1925), Simon (1955), Price (1976), Barabási & Albert (1999)
  - Rast miest, citácie vedeckých článkov, WWW,...

Preferential attachment (PA)

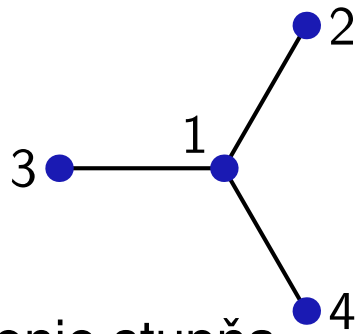
- Klasický model komplexnej siete
  - Yule (1925), Simon (1955), Price (1976), Barabási & Albert (1999)
  - Rast miest, citácie vedeckých článkov, WWW,...

- Postupne pridávame vrcholy a hrany (linky)
- Pravdepodobnosť, že vrchol získa nový link je úmerná stupňu tohto vrchola: $P(i,t) \sim k_i(t) + A$

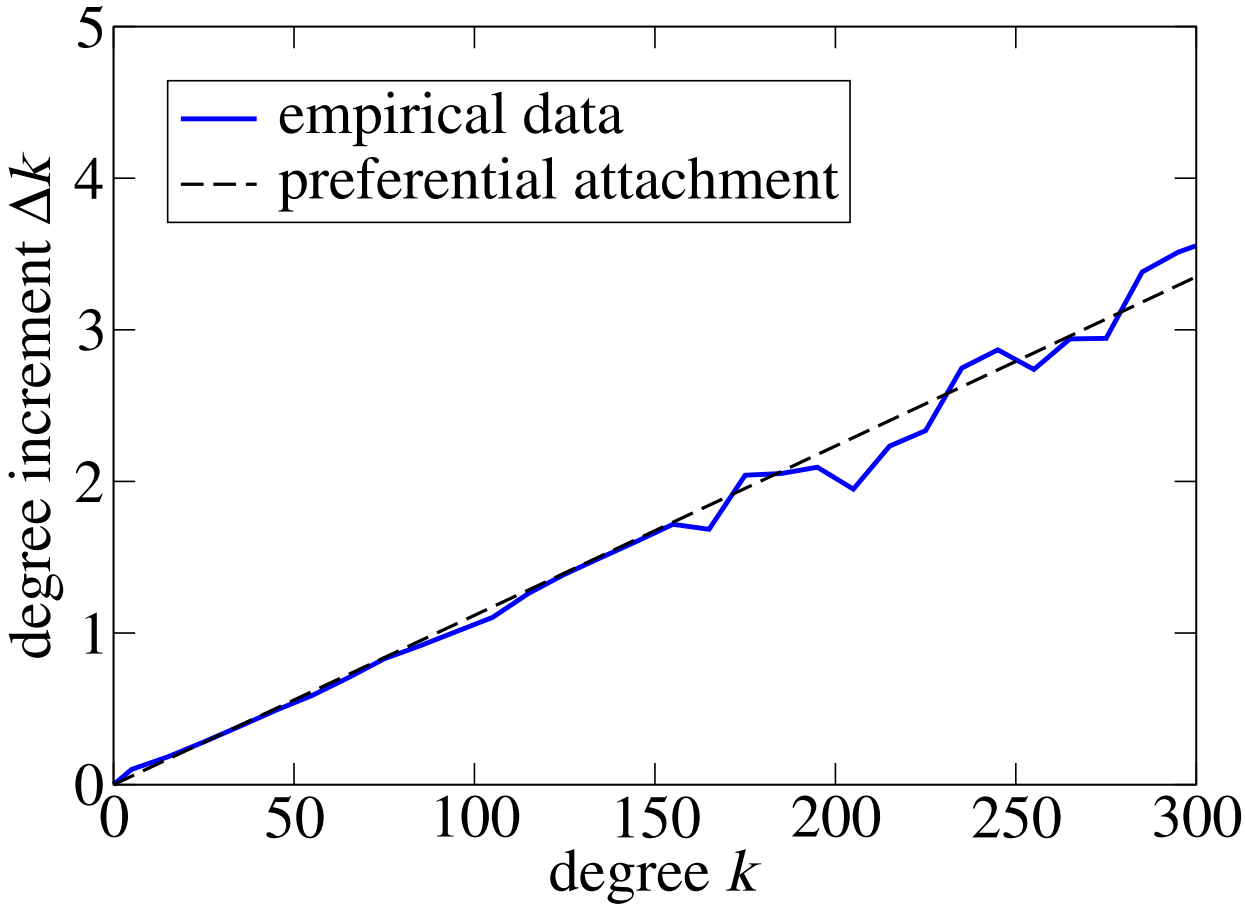


Preferential attachment (PA)

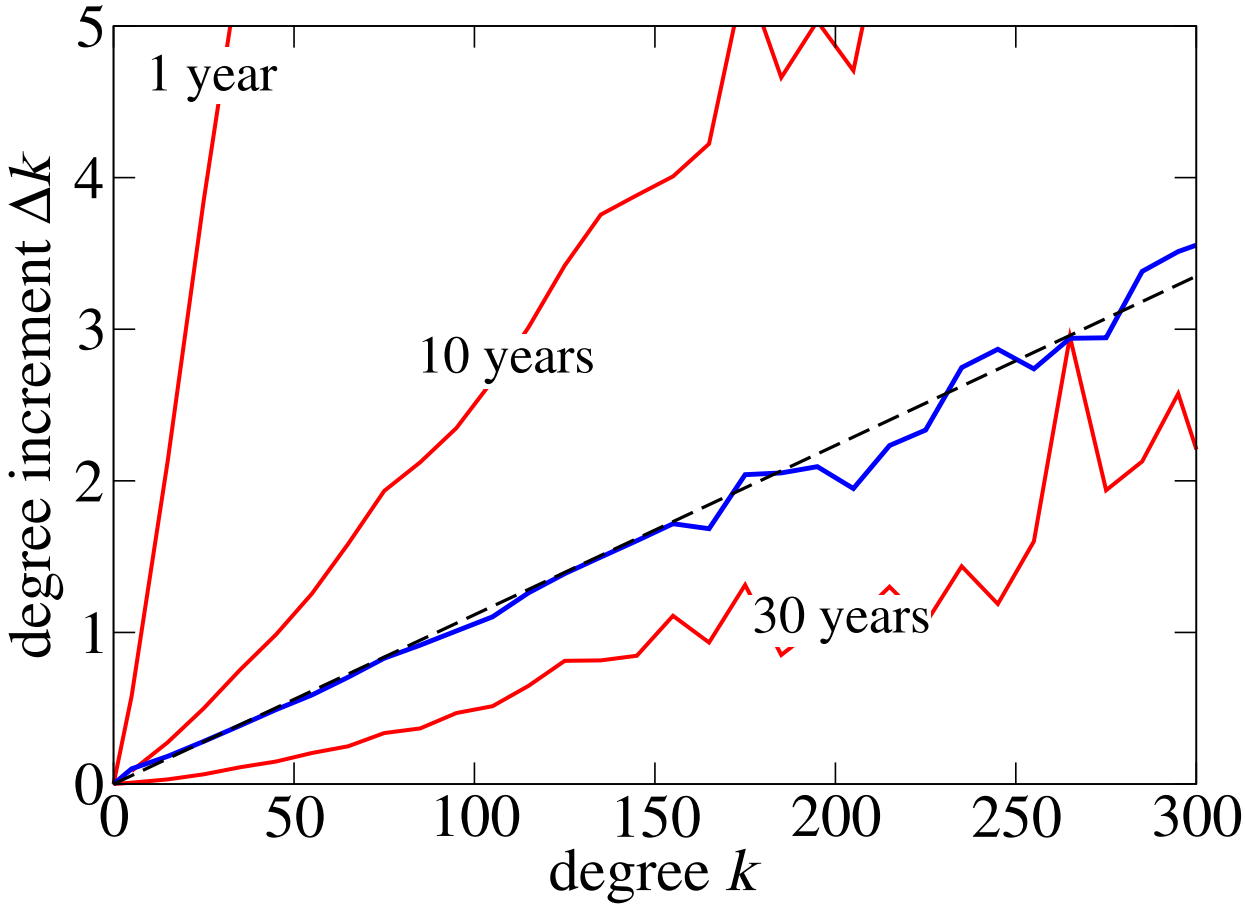
- Klasický model komplexnej siete
  - Yule (1925), Simon (1955), Price (1976), Barabási & Albert (1999)
  - Rast miest, citácie vedeckých článkov, WWW,...
- Postupne pridávame vrcholy a hrany (linky)
- Pravdepodobnosť, že vrchol získa nový link je úmerná stupňu tohto vrchola: $P(i,t) \sim k_i(t) + A$
  - Plusy: jednoduchosť, výsledné mocninové rozdelenie stupňa vrcholov
  - Mínusy: jednoduchosť (odchýlky od modelu pozorované v reálnych systémoch)



PA v citačných dátach



PA v citačných dátach



Pozri aj Adamic & Huberman (2000), Redner (2005), Newman (2009),. . .

Kľúčovú úlohu zohráva vek

"All the News That's Fit to Print."

The New York Times.

THE WEATHER.

TITANIC SINKS FOUR HOURS AFTER HITTING ICEBERG;
866 RESCUED BY CARPATHIA, PROBABLY 1250 PERISH;
ISMAY SAFE, MRS. ASTOR MAYBE, NOTED NAMES MISSING

Col. Astor and Bride, Isidor Straus and Wife, and Maj. Butt Aboard.

"RULE OF SEA" FOLLOWED

PICKS UP AFTER 8 HOURS

FRANKLIN HOPEFUL ALL DAY

HEAD OF THE LINE ABOARD

Biggest Liner Plunges to the Bottom at 2:20 A.M.

RESCUERS THERE TOO LATE

WOMEN AND CHILDREN FIRST

SEA SEARCH FOR OTHERS

Dve zovšeobecnenia základného PA

- Fitness model (Bianconi & Barabási, 2001):
  - Každý vrchol má svoju fitness, ktorá ovplyvňuje jeho atraktívnosť

$$P(i,t) \sim f_i k_i(t)$$

Dve zovšeobecnenia základného PA

- Fitness model (Bianconi & Barabási, 2001):
  - Každý vrchol má svoju fitness, ktorá ovplyvňuje jeho atraktívnosť

$$P(i,t) \sim f_i k_i(t)$$

- Starnutie vrcholov (Dorogovtsev & Mendes, 2000):
  - Ak sa vrchol objavil v čase s, pravdepodobnosť získania novej hrany v čase t je

$$P(i,t) \sim k_i(t)/(t-\tau_i)^{\alpha}$$

- Oba majú svoje problémy...

Nový model (PRL **107**, 238701, 2011)

1. Skombinujeme fitness a starnutie
   - Fitness so starnutím = relevancia

$$P(i,t) \sim R_i(t)k_i(t)$$

2. Vrcholy sú navzájom rôzne!
   - Napríklad štartovacie hodnoty $R_i(0)$ môžu byť volené náhodne

Riešime model

$$P(i,t) = \frac{k_i(t)R_i(t)}{\sum_{j=1}^{t} k_j(t)R_j(t)} = \frac{k_i(t)R_i(t)}{\Omega(t)}$$

Riešime model

$$\frac{\mathrm{d}\langle k_i(t)\rangle}{\mathrm{d}t} \approx P(i,t) = \frac{k_i(t)R_i(t)}{\sum_{j=1}^{t} k_j(t)R_j(t)} = \frac{k_i(t)R_i(t)}{\Omega(t) \approx \Omega^*}$$

Riešime model

$$\frac{\mathrm{d}\langle k_i(t)\rangle}{\mathrm{d}t} \approx P(i,t) = \frac{k_i(t)R_i(t)}{\sum_{j=1}^{t} k_j(t)R_j(t)} = \frac{k_i(t)R_i(t)}{\Omega(t) \approx \Omega^*}$$

$$\Downarrow$$

$$\left\langle k_i^F \right\rangle = \exp\left(\frac{1}{\Omega^*}\int_0^\infty R_i(t)\,\mathrm{d}t\right) = \exp\left(T_i/\Omega^*\right)$$

Riešime model

$$\frac{\mathrm{d}\langle k_i(t)\rangle}{\mathrm{d}t} \approx P(i,t) = \frac{k_i(t)R_i(t)}{\sum_{j=1}^{t} k_j(t)R_j(t)} = \frac{k_i(t)R_i(t)}{\Omega(t) \approx \Omega^*}$$

$$\Downarrow$$

$$\left\langle k_i^F \right\rangle = \exp\left(\frac{1}{\Omega^*}\int_0^\infty R_i(t)\,\mathrm{d}t\right) = \exp\left(T_i/\Omega^*\right)$$

- $R_i(t)$ nepodstatná, T_i rozhoduje
- Ω^* sa dá zvoliť tak, aby sme dosiahli taký $\langle k \rangle$ aké chceme

Rozdelenie stupňa vrcholov

- $k_i^F(T_i)$ má úzke rozdelenie $\Longrightarrow$ potrebujeme heterogénne vrcholy

$$\left\langle k_i^F(T_i) \right\rangle = \exp\left(T_i/\Omega^*\right)$$

Rozdelenie stupňa vrcholov

- $k_i^F(T_i)$ má úzke rozdelenie $\Longrightarrow$ potrebujeme heterogénne vrcholy

$$\left\langle k_i^F(T_i) \right\rangle = \exp\left(T_i/\Omega^*\right)$$

- Napríklad:
  1. $\varrho(T)$ gaussovské $\Longrightarrow$ log-normálne $P(k)$
  2. $\varrho(T)$ s exponenciálnym chvostom $\Longrightarrow$ $P(k)$ s mocninovým chvostom
  3. $\varrho(T) \sim \mathrm{e}^{-\alpha T}$ $\Longrightarrow$ $P(k) \sim k^{-3}$ (presne ako pri PA!)

Dáta

1. Citácie medzi článkami uverejnenými APS
2. Citácie medzi americkými patentmi
3. Užívateľské zbierky “bookmarkov”
4. Stiahnutia článkov z Econophysics Forum

data description	label	nodes	links	span/resolution	Δt
APS citations	APS	450k	4.7M	117 years/daily	91 days
U.S. patents	PAT	3.2M	24M	31 years/yearly	1 year
web bookmarks	WEB	2.3M	4.2M	4 years/daily	10 days
paper downloads	EF	600	16k	23 months/daily	10 days

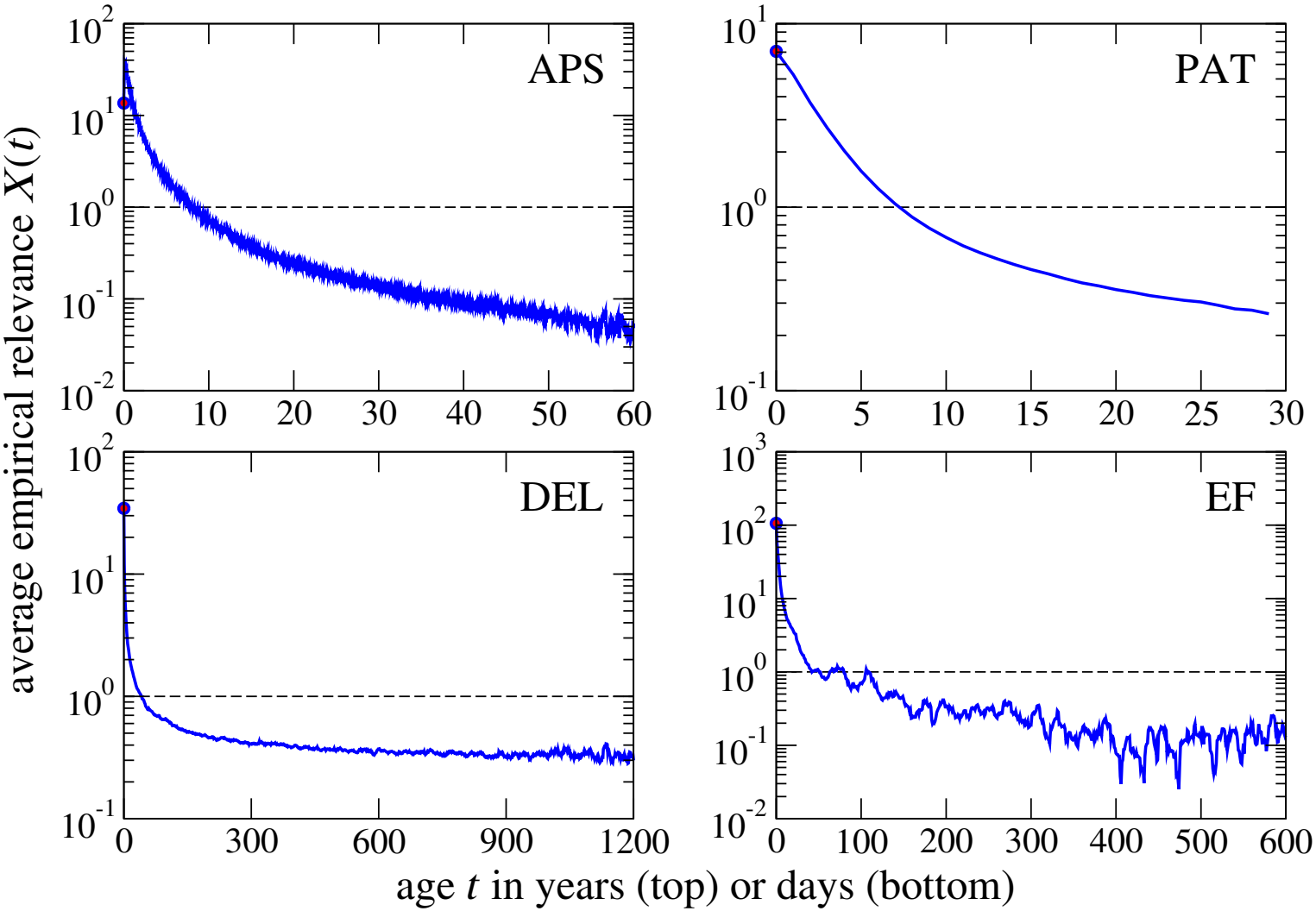
Empirická relevancia

- Empirická relevancia článku i v čase t: $X_i(t, \Delta t)$

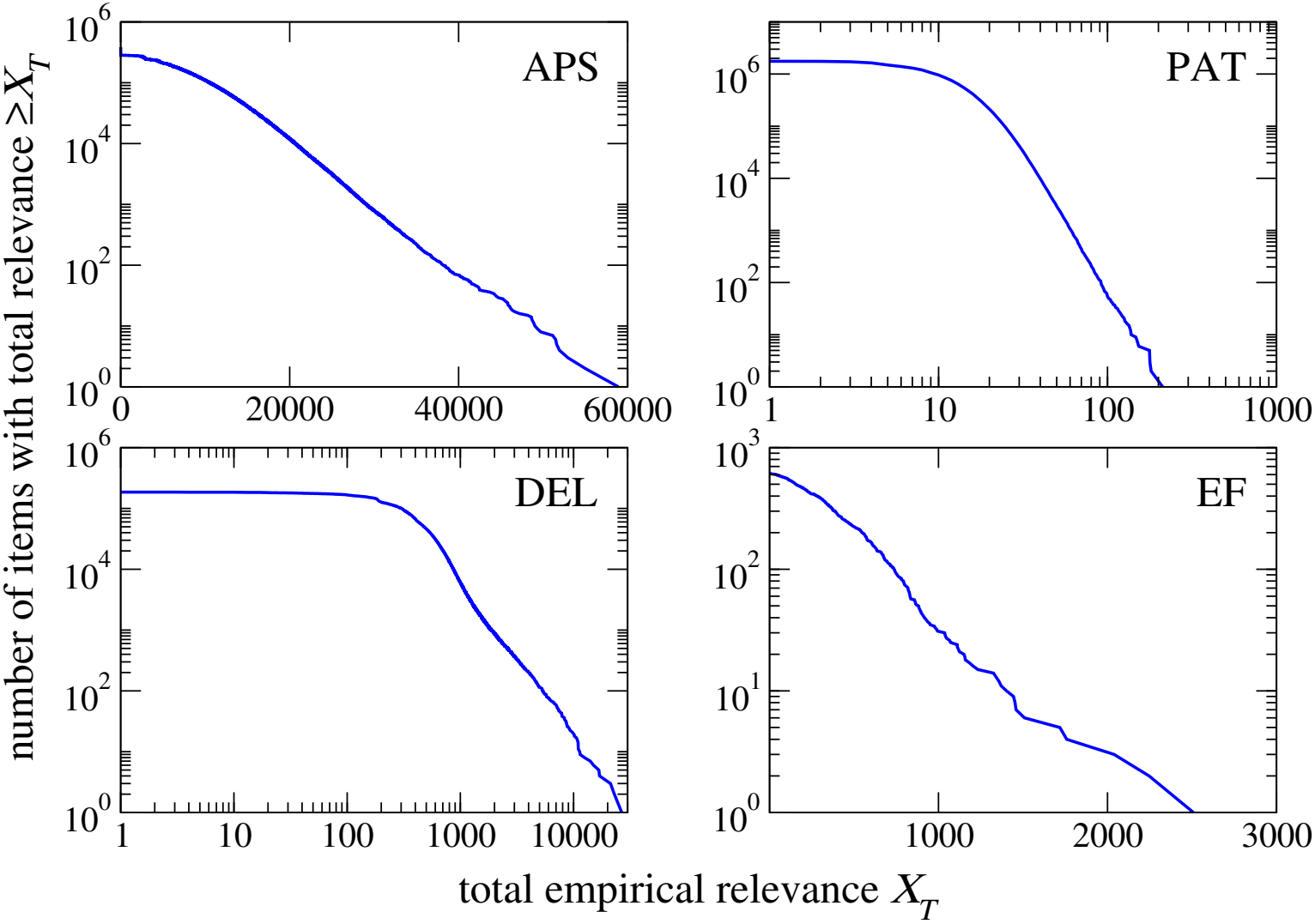
$$X_i(t, \Delta t) := \frac{\text{počet citácií článku } i \text{ počas } (t, t + \Delta t)}{\text{počet citácií očakávany podľa PA}} \sim \frac{\Delta k}{k(t)}$$

  - Keď PA pracuje ako má, $X_i(t, \Delta t) = 1$

Pokles relevancie



Heterogeneita relevancie



Aký dobrý je náš model?

- Potrebujeme robustné štatistické metódy

Aký dobrý je náš model?

- Potrebujeme robustné štatistické metódy
- Maximum likelihood: maximizujeme $\mathcal{L}(\mathcal{D}|M)$
  - $\mathcal{D}$ sú dané dáta (rastúca sieť)
  - M je parametrizovaný model
  - $\mathcal{L}$ je likelihood pre dané dáta vzhľadom na skúmaný model

Aký dobrý je náš model?

- Potrebujeme robustné štatistické metódy
- Maximum likelihood: maximizujeme $\mathcal{L}(\mathcal{D}|M)$
  - $\mathcal{D}$ sú dané dáta (rastúca sieť)
  - M je parametrizovaný model
  - $\mathcal{L}$ je likelihood pre dané dáta vzhľadom na skúmaný model
- Problémy:
  - **Dimenzionalita:** počet parametrov je úmerný počtu vrcholov
  - **Veľkosť dát:** výpočet likelihood je pomalý
  - **Konvergencia:** nič moc (plytké maximum)

Modely súťažia

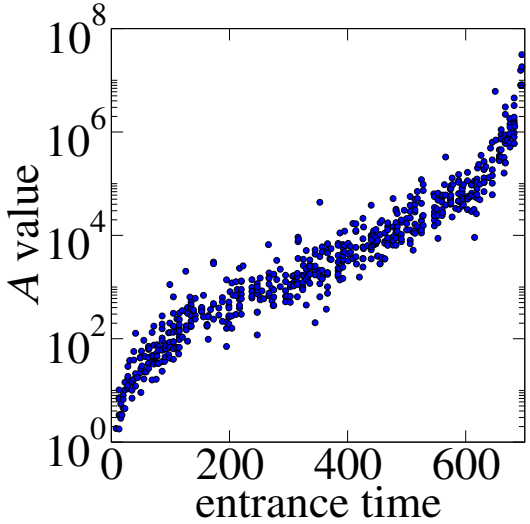
0. Hrany sa priradzujú k vrcholom náhodne
1. PA s aditívnym členom: $P(i,t) \sim k_i(t) + A$
2. PA s heterogénnym additívnym členom: $P(i,t) \sim k_i(t) + A_i$
3. Eom-Fortunato model: $P(i,t) \sim k_i(t) + A_i(t)$
4. PA s relevanciou: $P(i,t) \sim R_i(t)\big(k_i(t) + A\big)$

MLE pre dáta z Econophysics Forum

model	$\max \mathcal{L}(\mathcal{D}\|M)$	parametre
M_0	-5.55 ± 0.00	0
M_1	-5.52 ± 0.01	1
M_2	-4.44 ± 0.01	N
M_3	-4.00 ± 0.01	$N+3$
M_4	-3.94 ± 0.01	$N+4$

MLE pre dáta z Econophysics Forum

model	max $\mathcal{L}(\mathcal{D}\|M)$	parametre
M_0	-5.55 ± 0.00	0
M_1	-5.52 ± 0.01	1
M_2	-4.44 ± 0.01	N
M_3	-4.00 ± 0.01	$N+3$
M_4	-3.94 ± 0.01	$N+4$



- Čisté PA (M_1) skoro také zlé ako benchmark model M_0
- Heterogénny aditívny člen (M_3) má chabý zmysel
- Rozdiel M_4 vs M_3 vyzerá malý ale zodpovedá $\mathcal{D}$ 10^{410}-krát pravdepodobnejšie pri M_4 než pri M_3

Otvorené otázky

- Počítať ostatné vlastnosti siete (nie iba $P(k)$)
- Čo keď sa rast siete zrýchľuje?
- $\Omega(t)$ bez stacionárnej hodnoty

Otvorené otázky

- Počítať ostatné vlastnosti siete (nie iba $P(k)$)
- Čo keď sa rast siete zrýchľuje?
- $\Omega(t)$ bez stacionárnej hodnoty

- Problémy s konvergenciou MLE pri zvyšných dátach
- Ak poznáme dynamiku systému, môžeme začať predpovedať!

Ďakujem za pozornosť!

Otázky?