# Izboljšani populacijski modeli

Miha Čančula

11. januar 2012

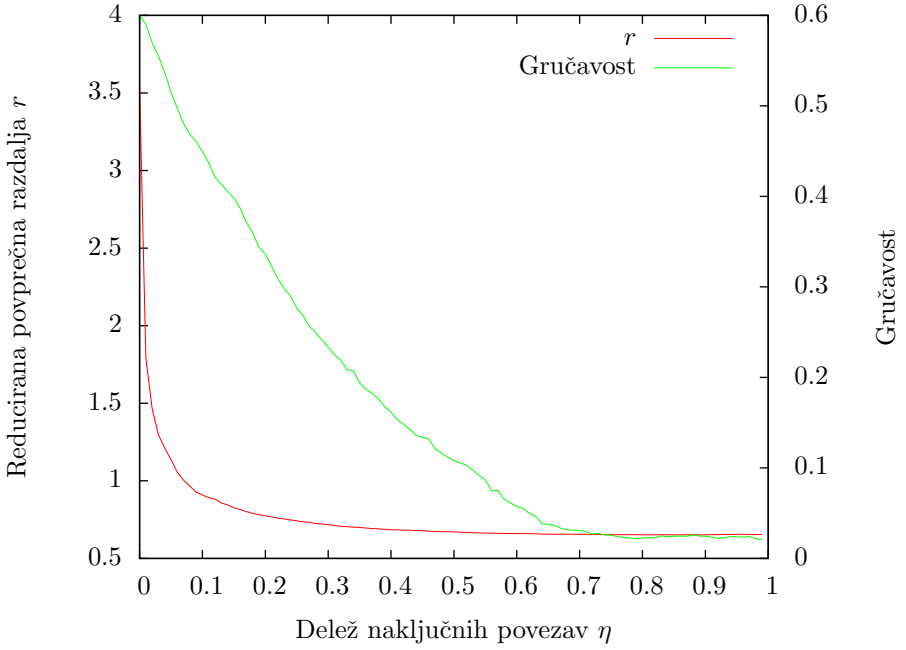
## 1 Model mali svet

### 1.1 Postopek

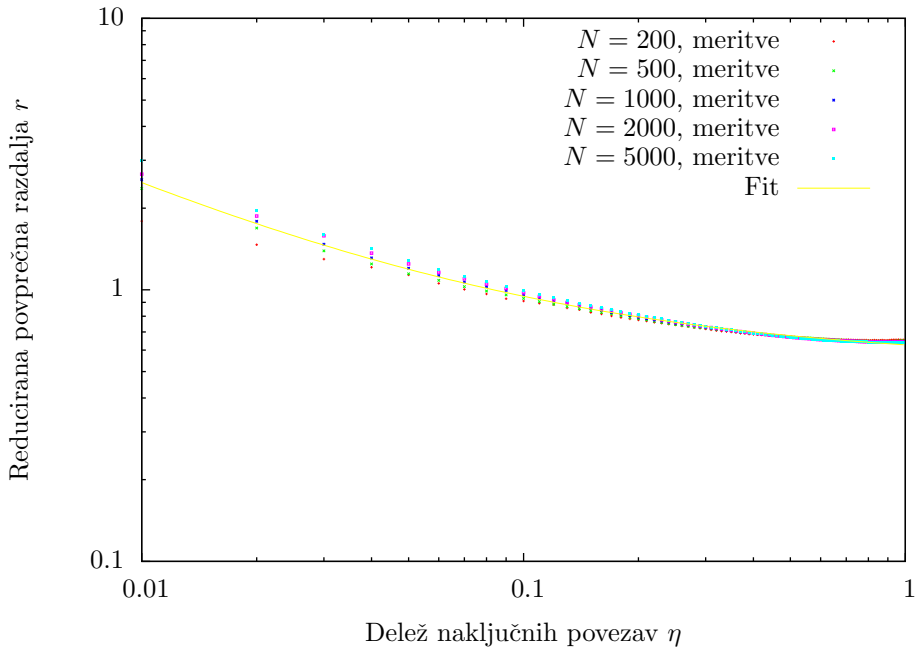
Ustvaril sem seznam $N$ elementov, ki so med seboj povezani sosedsko in ciklično, tako da je element $i$ povezana z elementi $\{(i+d) \mod N;\ |d| \leq k\}$. Povezanost $k$ sem za primere simulacije postavil na 3. Shranil sem seznam vseh prvotnih povezav.

Ob vsakem koraku sem eno izmed začetnih povezav nadomestil s povezavo med dvema naključnima elementoma. Poskrbel sem, da ta povezave prej ni obstajala, in da noben element ni povezan s samim seboj. Dodatno sem uvedel pogoj, da mora graf vseskozi ostati povezan, torej da obstaja povezava med poljubnima dvema točkama. Le za takšne grafe je namreč definirana povprečna razdalja, in le po njih se lahko širijo informacije oz. bolezni.

Ker sta izračuna povprečne razdalje med točkami in gručavosti grafa dolgotrajna postopka, ju nisem izvedel na vsakem koraku, ampak tako, da sta bila izvedena le po 100-krat.



Z grafa vidimo, da povprečna razdalja že pri majhnem številu dolgih povezav močno pade, nato pa se ustali pri neki konstantni vrednosti. Po drugi strani pa gručavost pada počasneje, na prvi pogled eksponentno.




Slika 1: Reducirana povprečna razdalja

Gručavost, definirana kot v navodilih, je brezdimenzijska količina med 0 in 1 in ni odvisna od velikosti grafa. To potrjuje graf na sliki 2.

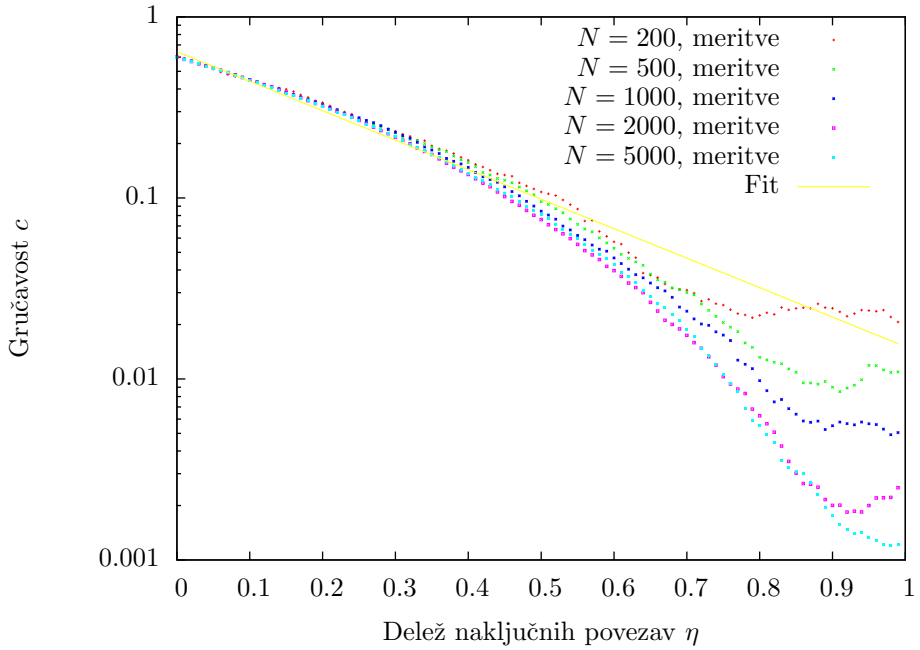
V naključnem grafu je povprečna razdalja med točkami sorazmerna z $\log N / \log k$, kjer je $k$ povprečna stopnja točke [1]. Za boljšo primerjavo sem torej namesto povprečne razdalje $\overline{d}$ računal z reducirano količino $r = \overline{d} \log k / \log N$. Ujemanje med krivuljami na grafu potrjuje to logaritemsko odvisnost.

Z grafov 1 in 2 se vidi, da reducirana povprečna razdalja in gručavost grafa nista odvisni od velikosti grafa, ampak le od razmerja med sosedskimi in naključnimi povezavami.

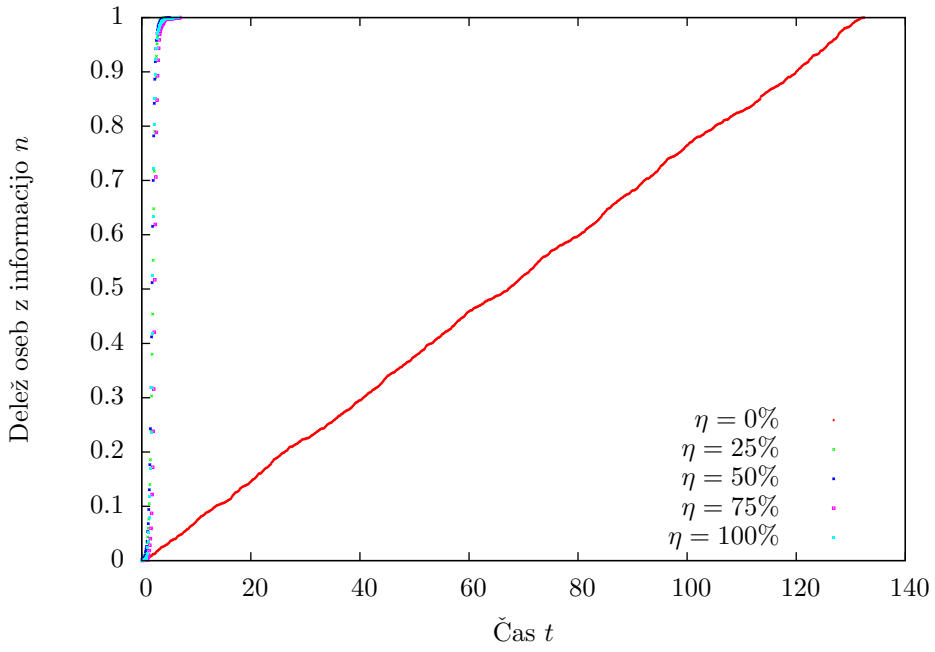
## 2 Širjenje informacije

Pri različnih deležih naključnih povezav $\eta$ sem simuliral širjenje informacije po grafu. Začel sem z enim osebkom, ki ima to informacijo, in jo v vsakem koraku z verjetnostjo $p$ razširi vsem sosedom, ki je še poznajo. Ta verjetnost je seveda sorazmerna z dolžino časovnega koraka $p \propto \Delta t$, zato sem ju v svojih brezdimenzijskih enotah kar izenačil.

Pri povsem sosedsko povezanem grafu je širjenje informacij linearen pojav. To lahko predpostavimo zaradi topologije grafa, informacija se širi po krožnici, ob vsakem koraku pa jo izvejo največ trije sosedje na vsaki strani. Že majhen prispevek naključnih povezav pa to dinamiko spremeni, in namesto linearne




Slika 2: Gručavost



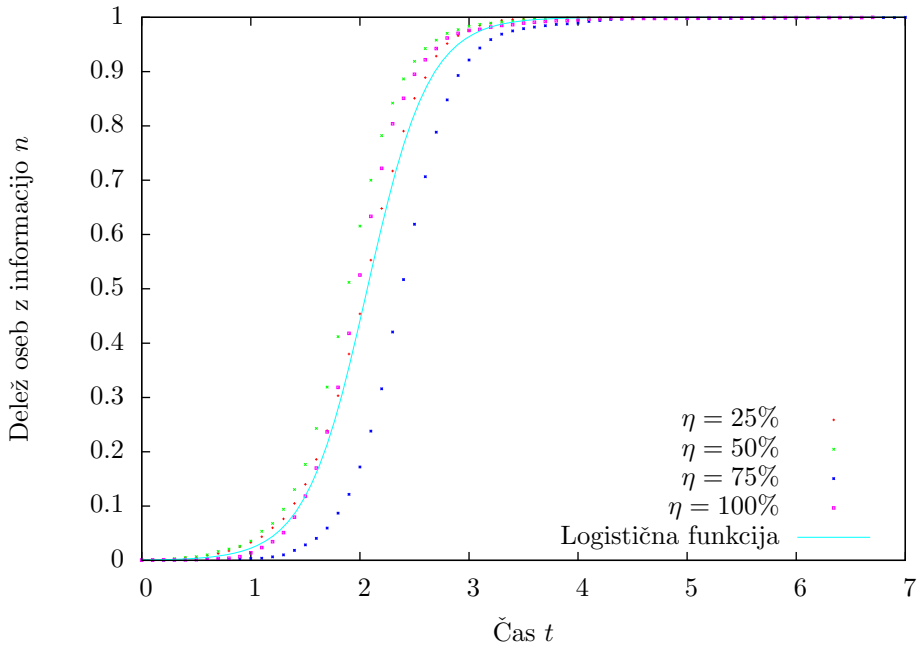

Slika 3: Širjenje informacije po grafu, $N = 2000$

časovne odvisnosti dobimo znano logistično enačbo. Oblika in položaj krivulje se z nadaljnjim večanje $\eta$ le malo spreminja, zato sem lahko vse štiri primere približal z eno samo krivuljo

$$n(t) = \frac{1}{1 + e^{-a(t-t_0)}} \tag{1}$$

$$t_0 = 2,07 \pm 0,01 \tag{2}$$

$$a = 3,5 \pm 0,1 \tag{3}$$

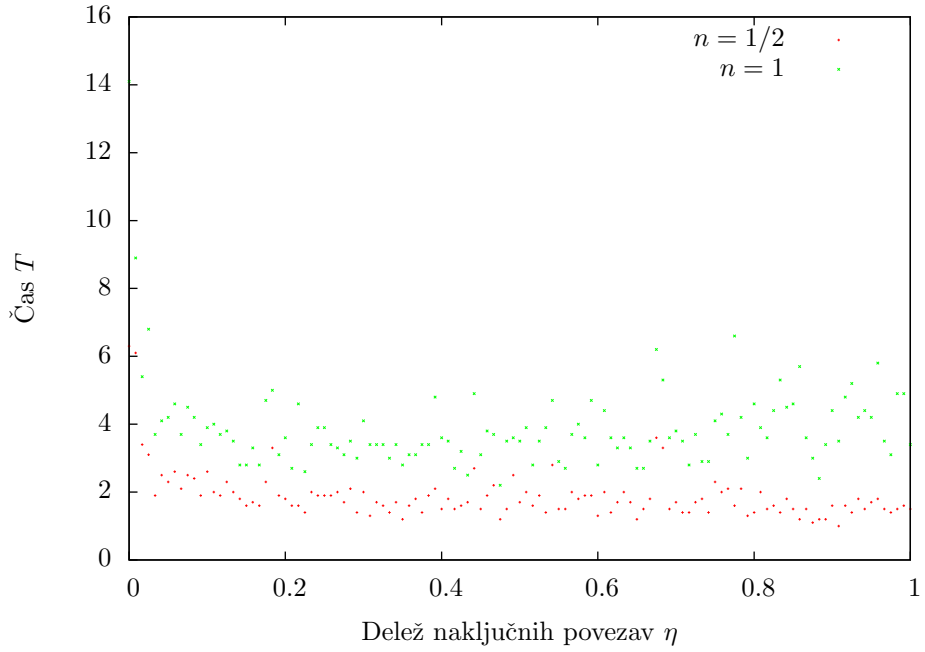



Slika 4: Širjenje informacije po grafu, $N = 2000$

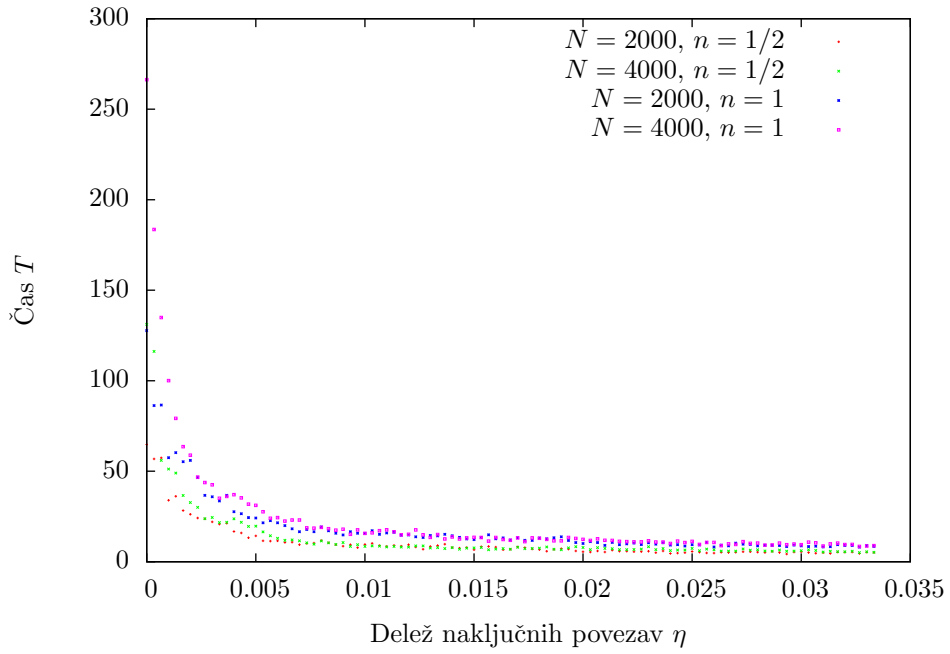
Ogledal sem si, kako hitro dosežemo, da polovica osebkov pozna informacijo ($n = 1/2$) pri različnih deležih naključnih povezav.

V vseh primerih je čas od polovične do polne informiranosti skoraj enak, zato je dovolj če obravnavamo le enega. Že pri uvedbi majhnega števila (manj kot 10%) naključnih povezav dolgega dosega čas razširjanja močno pade, pri $\eta = 20\%$ pa se ustali in nadaljnja odstopanja lahko pripišemo statističnemu šumu. Najzanimivejša je odvisnost pri majhnih $\eta$, zato sem se temu področju posvetil bolj podrobno z večjimi grafi.

Za večje grafe se čas širjenja informacije ustali že prej, pri okrog 1% naključnih povezav. To pomeni, da potrebujemo le majhno število povezav dolgega dosega, da zagotovimo hitro širjenje informaci.



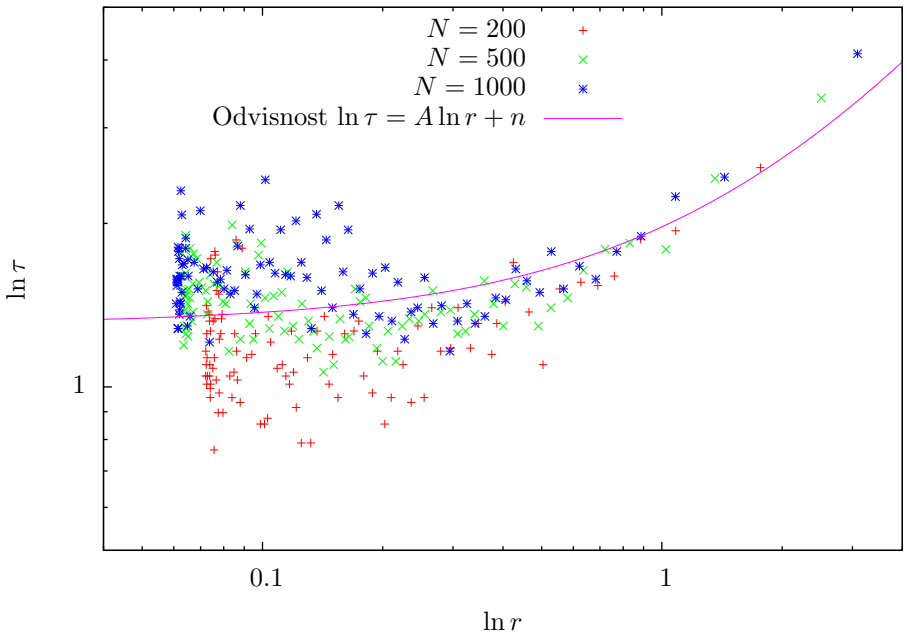

Slika 5: Hitrost širjenja informacije po grafu, $N = 200$




Slika 6: Čas širjenja informacije po grafu z večjim številom točk

## 2.1 Odvisnost od povprečne razdalje

Čas širjenja informacije ima na prvi pogled podobno odvisnost od $\eta$ kot povprečna razdalja med točkami, zato sem ju primerjal med seboj. Ker obe količini zelo hitro padeta proti neki končni vrednosti, sem vsako izmed njih dvakrat logaritmirav.




Slika 7: Odvisnost med povprečno razdaljo in časom, potrebnim za polovično razširitev informacije. Na obeh oseh sta logaritma količin, pa tudi sam graf je v logaritemskem merilu, tako da je to dvojno-logaritemski graf za obe količini.

Odvisnost dokaj dobro popiše izraz

$$\ln\tau = k\ln r + n = \ln\left(r^k\right) + n \tag{4}$$

$$\tau = \exp\left(\ln r^k + n\right) = e^n \cdot r^k = Cr^k \tag{5}$$

# 3 Populacijski modeli

Ker imamo velikost populacije dosti bolj natančno podano v letih okrog 2000, meritve oz. ocene pa segajo do pred nekaj milijoni let, sem moral uporabiti logaritemsko časovno skalo. Za to borajo biti vse vrednosti pozitivne, najbolje pa je, če je vrednost 0 v bližini leta 2000, kjer imamo največ meritev. Zato sem časovno skalo kalibriral z brezdimenzijsko spremenljivko

$$x(t) = 2202 - \frac{t}{1\text{ leto}} \tag{6}$$

Konstanto 2202 sem izbral, ker je zgornja meja za meritve, ocene in predvidevanja v tabeli. Podatki segajo do leta 2200, z izbiro vrednosti 2202 sem zagotovil, da obstaja tudi drugi logaritem, torej da je $\ln(x) > 0$.

## 3.1 Model Kapice

Diferencialno enačbo Kapice

$$\dot{n} = K \sin^2 \frac{n}{K} + \frac{1}{K} \tag{7}$$

lahko integriramo in dobimo [2]

$$n(t) = K \arctan \left[ \frac{\tan\left(\frac{(t-t_1)\sqrt{K^2+1}}{K^2}\right)}{\sqrt{K^2+1}} \right] \approx K \arctan \left[ \frac{\tan \frac{t-t_1}{K}}{K} \right] \tag{8}$$

$$N(T) = K^2 \arctan \left[ \frac{\tan \frac{T-T_0}{\tau K}}{K} \right] \tag{9}$$

To velja ob predpostavki, da $N(T_0) = 0$, torej za $T_0$ postavimo začetek človeštva, $T_0 \approx -4,4 \cdot 10^6$ let. Konstanti $K$ in $\tau$ ocenimo iz podatkov, primerni vrednosti sta $K = 67000$ in $\tau = 42$ let. Obe imata tudi praktičen pomen, $K$ je minimalna velikost stabilne in samozadostne civilizacije, $\tau$ pa je blizu povprečnega življenjskega časa enega človeka.

Paziti moramo tudi, za bo funkcija zvezna in vedno pozitivna, tudi ko ima tangens pol. Takrat je tudi rast populacije najhitrejša. Čas, ob katerem se to zgodi, označimo s $T_2$.

$$\frac{T_2 - T_0}{\tau K} = \frac{\pi}{2} \tag{10}$$

$$T_2 = T_0 + \frac{\tau K \pi}{2} \approx 0 \tag{11}$$

Oba člena na desni imata podoben velikostni red, torej do prehoda pride nekje v bližini našega stetja. V modelu, ki ga prilagajam podatkov, namesto časa $T_0$ raje uporabil $T_2$, saj je bližje zanimivemu delu odvisnosti, pa tudi večini meritev.

Inverzni tangens v enačbi uspešno odpravi neskončnost v polu, zaradi predznaka in zveznosti krivulje pa moramo izrazu pri $T > T_2$ prišteti še $\pi K^2$.

$$N(T) = K^2 \left( \arctan \left[ \frac{\tan\left(\frac{T-T_1}{\tau K} + \frac{\pi}{2}\right)}{K} \right] + \pi \Theta(T - T_1) \right) \tag{12}$$

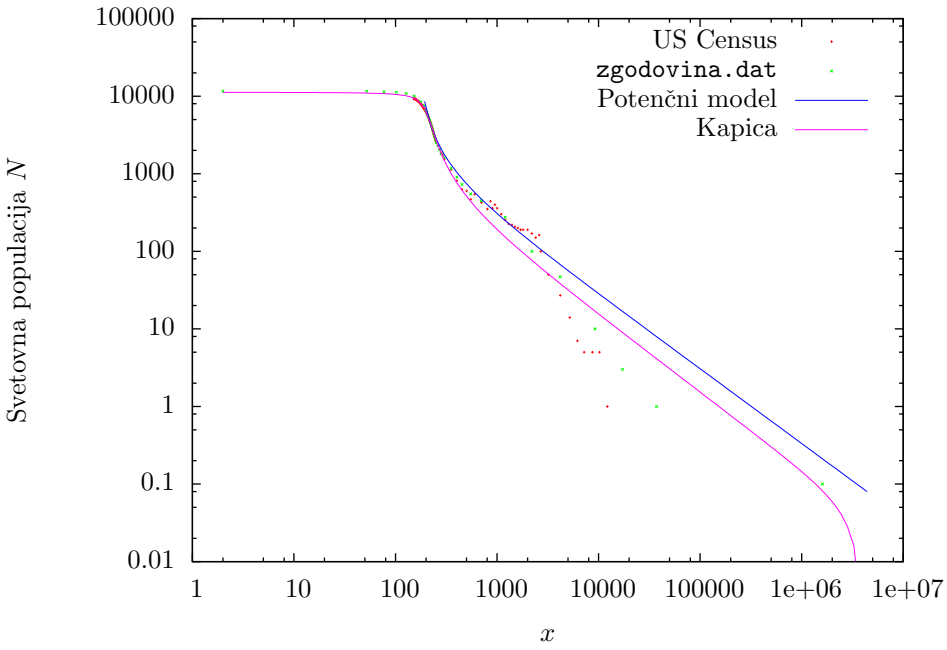
## 3.2 Podatki

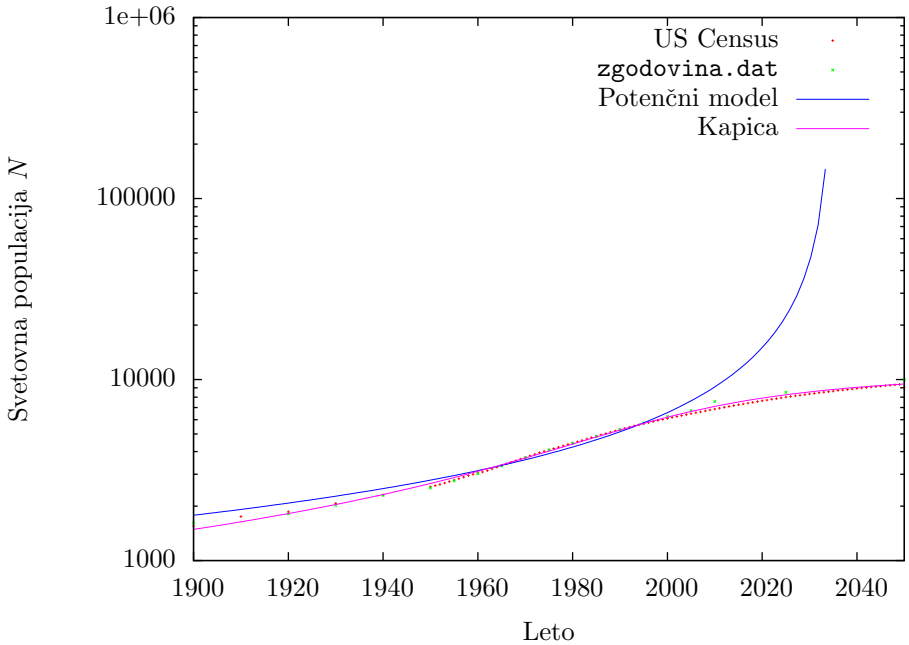
Uporabil sem dva nabora podatkov:

1. Datoteka `zgodovina.dat`, dobljena s strani predmeta

2. Podatke US Census Bureau [3].

Vsi podatki so seveda zgolj ocene, tisti iz prihodnosti pa napovedi. Ker so napovedi narejene pa podlagi istih modelov, kot jih trenutno poskušam prilagoditi, sem pri fitanju uporabil le podatka izpred leta 2000.




Prilagoditvi obeh modelov podatkov ameriškega urada nam data podatke v tabelah 1 in 2.

| Parameter | Vrednost |
|---|---|
| $C$ | $(2,0 \pm 0,5) \cdot 10^{11}$ |
| $\alpha$ | $-0,96 \pm 0,05$ |
| $T_1$ | $2035 \pm 4$ |
| $\chi^2_{red}$ | 24000 |

Tabela 1: Optimalni parametri za potenčni model

| Parameter | Vrednost |
|---|---|
| $K$ | $(62 \pm 1) \cdot 10^3$ |
| $\tau$ | $40,3 \pm 0,5$ |
| $T_2$ | $1998 \pm 2$ |
| $\chi^2_{red}$ | 6200 |

Tabela 2: Optimalni parametri za model Kapice

# 4 Velikosti in vrsti red

## 4.1 Izpeljava

Izraz

$$\frac{U(R)}{U_0} = \frac{\ln U_0}{R + \ln U_0} \tag{13}$$

$$\tag{14}$$

lahko obrnemo tako, da izrazimo $R(U)$

$$\frac{\ln U_0}{R + \ln U_0} = \frac{U(R)}{U_0} \tag{15}$$

$$\frac{R + \ln U_0}{\ln U_0} = \frac{U_0}{U(R)} \tag{16}$$

$$R(U) = \left(\frac{U_0}{U} - 1\right) \ln U_0 \tag{17}$$

Če želimo porazdelitev mest po velikosti, moramo izraz odvajati. Število mest z velikostjo med $U_1$ in $U_2$ je nareč kar enako $R(U_1) - R(U_2)$. Z limitiranjem $U_2 \to U_1$ dobimo izraz za diferencialno verjetnostno gostoto

$$p(U) = \left|\frac{\partial R}{\partial U}\right|_U = \frac{U_0 \ln U_0}{U^2} \propto U^{-2} \tag{18}$$

Takšna odvisnost torej popiše primere, kjer pogostost pojavljanja pada z drugo potenco velikosti.

Iz izraza (13) pa lahko izračunamo tudi skupno število prebivalcev v mestih:

$$N = \sum_{R=1}^{M} U(R) \approx \int_1^\infty U(R)\,\mathrm{d}R = U_0 \ln U_0 \int_1^M \frac{\mathrm{d}R}{R + \ln U_0} \tag{19}$$

$$= U_0 \ln U_0 \left[\ln(R + \ln U_0)\right]_1^M = U_0 \ln U_0 \ln\left(\frac{M + \ln U_0}{1 + \ln U_0}\right) \tag{20}$$

Število mest $M$ lahko ocenimo tako, da postavimo spodnjo mejo za njihovo velikost $U_m$. Tedaj je število mest kar enako rangu najmanjšega mesta $R(U_m)$ in je odvisno le od $U_0$. V sistemu, ki ga opisuje takšen model, torej obstaja direktna povezava med skupnim število prebivalcev mest $N$ in velikostjo največjega mesta $U(1)$, saj sta oba odvisna le od enega parametra $U_0$.

V primerih, ko ta model ni zadovoljivo opisal dejanskega stanja, sem uporabil bolj splošen izraz

$$U(R) = U_0 \frac{a}{x + a} \tag{21}$$

Tak model opiše verjetnostno porazdelitev po velikosti

$$R(U) = a \left( \frac{U_0}{U} - 1 \right) \tag{22}$$

$$\frac{\partial R}{\partial U} = \frac{a U_0}{U^2} \tag{23}$$

Verjetnostna gostota tudi sedaj pada s kvadratom velikosti, v modelu pa nastopata dva prosta parametra. Zato sta lahko skupno število prebivalcev in velikost največjega mesta izberemo poljubno. Tak model dobro opiše tudi nekatere odvisnosti, ki jih prvi ne.

## 4.2 Ocena napake

Napako velikosti vsakega mesta sem ocenil na $\sqrt{U}$, kjer je $U$ število prebivalcev. Ta napaka ne pride nujno iz meritev, saj so štetja prebivalstva tako v Sloveniji kot v Ameriki pogosta in točna, ampak iz stalnega preseljevanja prebivalcev in nejasnih meja mest. Ker sem vse rezultate predstavljal z logaritemskimi grafih, je na ta način vizualno ujemanje na grafu boljše kot pri konstanti napaki.

## 4.3 Mesta

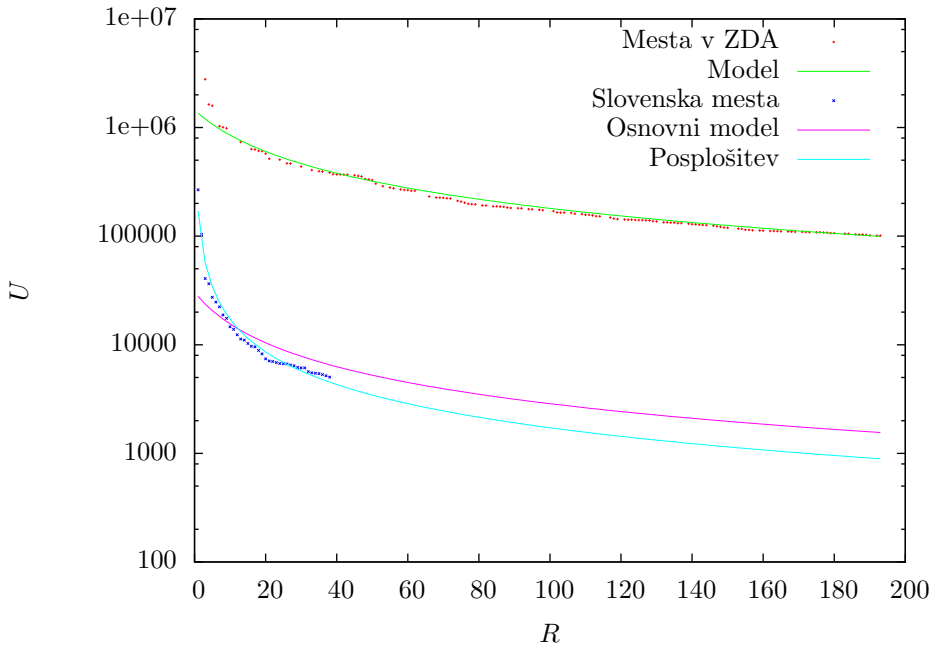
Model sem poskušal prilagoditi podatkom za mesta v Sloveniji in v Združenih državah Amerike. Rezultati so prikazani na sliki 8.

Model res dobro opiše velikosti mest v ZDA. Odstopanja so vidna le pri treh največjih mestih, ki so večja kot bi pričakovali.

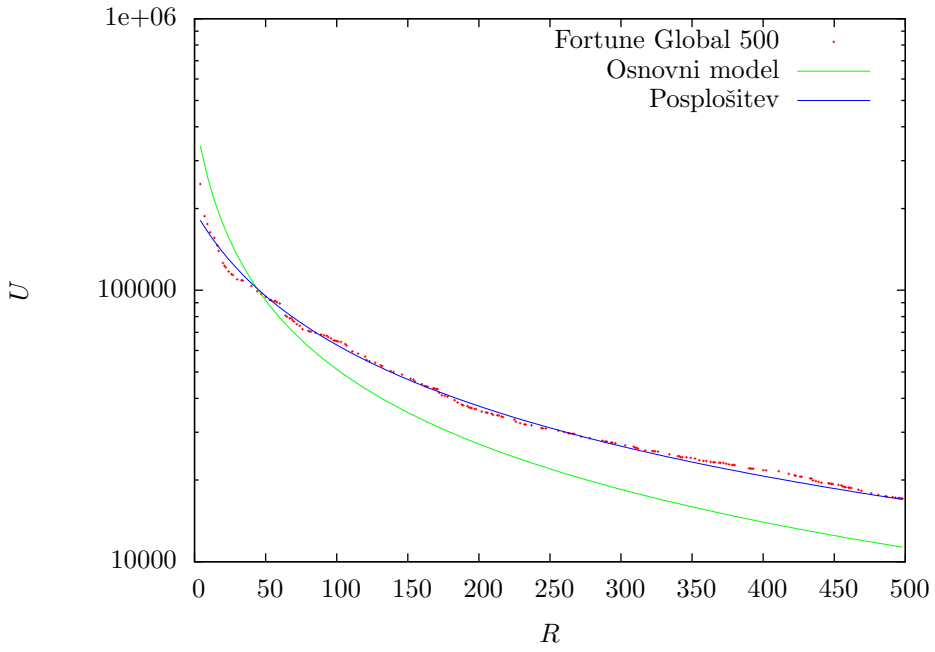
Za slovenska mesta pa model ni primeren, saj velikost mest pada dosti hitreje kot pričakovano, in to po celotni lestvici. Boljše ujemanje pa dobimo s posplošenim modelom, z močno zmanjšanim $a$. Optimalna vrednost za $a$ je namprec le 0.01, $U_0$ pa primerno velik, torej se sistem obnaša približno kot $U(R) \propto 1/R$.

## 4.4 Podjetja in internetne strani

Isti model sem preizkusil tudi na nekaterih drugih naborih podatkov. Izkaže se, da prihodki velikih podjetij [4] ne sledijo zgoraj porazdelitvi v izrazu (18). V nasprotju s Slovenskimi mesti pa je odstopanje tu v drugo smer: Največja podjetja imajo manjše prihodke, kot bi pričakovali. Optimalna vrednost za parameter $a$ je tu večja od $\ln U_0$.

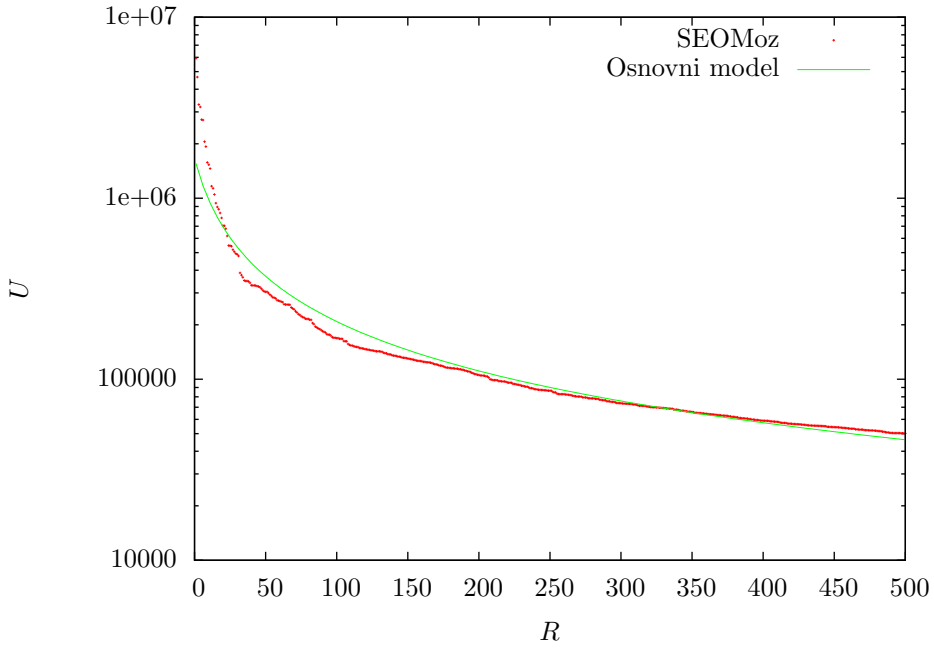



Slika 8: Velikosti mest v ZDA in v Sloveniji




Slika 9: Letni prihodki največjih svetovnih podjetij

Dosti bolje se izkažejo popularne spletne strani. Podatki o številu povezav na posamezne domene, ki jih objavlja podjetje SEOMoz [5], se dobro ujemajo z našim modelom.

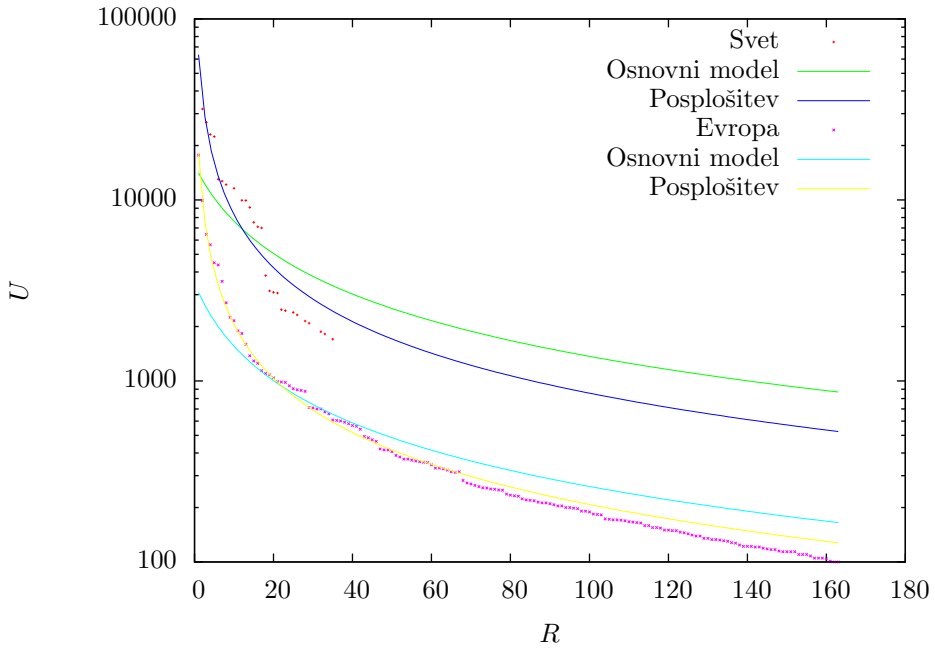



Slika 10: Število zunanjih povezav na največje svetovne spletne strani

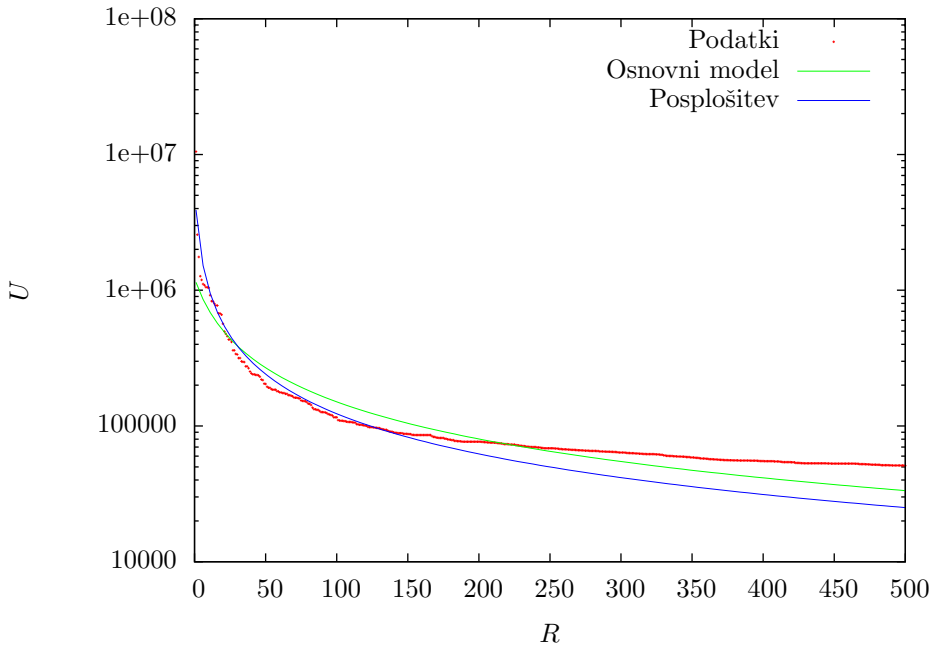
Preveril sem tudi, ali je opisana odvisnost značilna le za povezovanje ljudi v skupine, ali se pojavlja tudi v neživi naravi. V ta namen sem obravnaval velikosti največjih svetovnih in evropskih jezer [6, 7]. Podatki in približek z obema modeloma so na sliki 11

Osnovni model z enim parametrom tu ni primeren, s posplošitvijo pa dobro opiše realno stanje. Tako kot pri mestih v Sloveniji tu velikost prehitro pada z vrstnim številom, zato moramo parameter $a$ zmanjšati. Namesto pričakovane vrednosti okrog 12 je optimalen $a$ med 0,1 in 0,4.

Kot pravi računalniški fizik sem si ogledal tudi svoja najljubša orodja: superračunalnike. Narisal in obravnaval sem število operacij z realnimi, ki jih je računalnik zmožen opraviti na sekundo, tako imenovane FLOPS. Kot vidimo na sliki 12, že osnovni model dokaj dobro opiše njihovo porazdelitev, posplošitev z dodatnim parametrom pa ne prinese skoraj nobene koristi.




Slika 11: Površine največjih svetovnih jezer




Slika 12: Hitrost 500 najmočnejših superračunalnikov

# Literatura


[1] Fan Chung and Linyuan Lu, The average distances in random graphs with given expected degrees, Proceedings of the National Academy of Sciences of the United States of America, 2002.

[2] Denis Brojan, Populacijska dinamika človestva, seminar, 2009.

[3] U. S. Census Bureau, http://www.census.gov/population/international/data/idb/worldhis.php, http://www.census.gov/population/international/data/idb/worldpoptotal.php.

[4] Fortune Global 500, http://money.cnn.com/magazines/fortune/global500/2010/full_list/index.html.

[5] SEOMoz, http://www.seomoz.org/top500.

[6] http://www.factmonster.com/ipka/A0001777.html

[7] http://en.wikipedia.org/wiki/List_of_largest_lakes_of_Europe