Sveučilište u Zagrebu
Fakultet Elektrotehnike i Računarstva

Završni rad br. 123

Identifikacija ključnih igrača u kompleksnim mrežama blogova

Matija Piškorec

mentor: *Prof.dr.sc. Branko Jeren*

11. lipnja 2008.

Zahvale

Zahvaljujem prije svega Mili Šikiću i Hrvoju Štefančiću sa Instituta Ruđer Bošković u Zagrebu uz čiju je suradnju ovaj rad ostvaren.

Mnogi dijelovi programskog koda preuzeti su od kolega koji su ranije radili na sličnim projektima. Bez toga ovaj projekt sasvim sigurno ne bi bio završen na vrijeme. Zbog savjeta, sugestija i izdvojenog vremena, posebno hvala Ivoru Prebegu i Ivanu Milošu.

Sadržaj

Uvod

Kako definirati *centralnost*? U kolokvijalnom izričaju pojam je jednostavno objasniti - centralno je grijanje, banka koja ima monopol na izdavanje zakonskog sredstva plaćanja ili onaj igrač koji na nogometnom terenu igra centar. S druge strane, formalno definiranje centralnosti nije trivijalan problem jer prije svega ovisi o kontekstu u kojem se svojstvo centralnosti pojavljuje. Poopćeno gledano, centralnost pretpostavlja postojanje entiteta i veza između njih te kriterija po kojem će određeni entitet biti vrednovan kao centralan.

Formaliziranjem takvog modela dolazimo do pojma *mreže* i pripadne teorije koja se prvobitno razvijala kao grana matematike zbog čega je i poznatija pod nazivom *teorija grafova*. Ipak, otvaranjem cijelog područja novih problema u zadnjem desetljeću, mrežnu paradigmu preuzele su i brojne druge grane znanosti - fizika, biologija i sociologija između ostalih. [1]

U ovom radu ukratko će se predstaviti tri danas popularne mjere centralnosti - centralnost stupnja (poseban slučaj Sadeove k-path centralnosti), Freemanova međupoloženost i PageRank. Centralnost stupnja i PageRank implementirat će se i testirati na modelu mreže blogova gdje blogeri imaju ulogu čvorova a njihovi međusobni komentari ulogu veza. Podatci za modeliranje mreže uzeti su s blog servisa `www.bloger.hr` koji djeluje od 2004. i danas broji oko 13000 registriranih blogera.

Osim gore navedenih mjera temeljenih na strukturalnom pristupu u ovom radu bit će predložena i jedna nova mjera koja se temelji na lokalnim karakteristikama samog čvora. Te karakteristike, za razliku od primjerice stupnja čvora, moraju biti nevezane za lokalnu ili globalnu strukturu mreže i moraju dobro odgovarati konceptu *važnosti* ili *autoriteta* koji se često dovodi u vezu s centralnošću kod socijalnih mreža. Riječ je o *vanjskom odzivu na vrhovima* koji u obzir uzima karakteristike svih komentara i postova koji pripadaju danom blogeru. U nastavku teksta skraćeno će se govoriti samo o *odzivu*. Razvijeno je devet modela mjere odziva koji su implementirani i uspoređeni s PageRankom i centralnosti prema stupnju čvora. Provedena su rangiranja svih blogova prema pojedinim mjerama i rezultati su uspoređeni.

Jedan od ciljeva ovog istraživanja je provjeriti u kojoj mjeri razni modeli odziva koreliraju s ostalim mjerama centralnosti baziranih na strukturalnom pristupu.

[1] Upravo zbog toga je prikladnije koristiti termine mreža, čvor i veza (eng. *network*, *node*, *connection*) umjesto graf, vrh i brid (eng. *graph*, *vertice*, *edge*) koji su uobičajeni u teoriji grafova.

1 Strukturalna centralnost

Trenutno još uvijek ne postoji jedinstvena definicija *centralnosti čvora* i tom problemu se pristupa na različite načine. Čak i samo klasificiranje danas poznatih mjera centralnosti zahtjevan je zadatak te odabir pojedine mjere prije svega ovisi o kontekstu ili o njenoj praktičnoj primjenjivosti na određeni problem. Kako je svim mjerama zajedničko da na neki način opisuju ulogu čvora prilikom šetnje kroz mrežu Borgatti i Everett u [1] navode tri osnovne dimenzije prema kojima se vrednuju mjere centralnosti: *pozicija šetnje*, *tip šetnje* i *svojstva šetnje*.

> **Pozicija šetnje** (eng. *walk position*): Mjere se dijele na radijalne (eng. *radial measures*) i medijalne (eng. *medial measures*). Radijalne u obzir uzimaju samo šetnje koje započinju (ili završavaju) u pojedinom čvoru, a medijalne su sve one koje bar jednom prolaze kroz pojedini čvor.
>
> **Tip šetnje** (eng. *walk type*): Tip šetnje ovisi o ograničenjima koje se nameću šetnjama i koje se uzimaju u obzir. Tipična ograničenja su, na primjer, uzimanje u obzir samo najkraćih puteva između dva čvora (eng. *geodesic paths*), puteva koji nemaju zajedničkih veza (eng. *edge-disjoint paths*) ili puteva koji nemaju zajedničkih vrhova (eng. *vertex-disjoint paths*).
>
> **Svojstva šetnje** (eng. *walk property*): Svojstva koja se kod šetnji mogu uzimati u obzir su duljina (eng. *length*) ili njihov broj (eng. *volume*).

Ovisno o poziciji i svojstvu šetnje mjere centralnosti se mogu podijeliti u četiri disjunktne kategorije. Kako sugeriraju Borgatti i Everett u [1] usporedba mjera ima smisla samo između mjera iz iste skupine, dok je one iz različitih najbolje gledati kao komplementarne. U nastavku rada opisat će se dvije mjere iz različitih kategorija:

> **k-centralnost**: Gleda se broj šetnji koje izviru iz pojedinog čvora. Stoga je pozicija šetnje radijalna, a svojstvo šetnje volumen (broj).
> **Freemanova međupoloženost**: Gleda se broj šetnji koje prolaze kroz pojedini čvor. Stoga je pozicija šetnje medijalna, a svojstvo šetnje volumen (broj).

Mjera centralnosti PageRank, također je temeljena na strukturalnom pristupu, no zbog svoje specifičnosti ne može se kategorizirati prema gore navedenim kriterijima. Ipak, zbog svoje višestruko dokazane djelotvornosti (u teorijskoj i komercijalnoj domeni) uključena je u ovaj rad.

Mjere koje nisu temeljene na strukturalnom pristupu, također izbjegavaju klasifikaciju prema strukturalnim kriterijima. One u obzir uzimaju i informacije koje nisu vezane isključivo uz strukturu same mreže - informacije o svojstvima samih čvorova na primjer. Nekoliko modela jedne takve mjere bit će opisane u sljedećem poglavlju.

k-centralnost (*k-path centrality*)

Jedna od najjednostavnijih mjera centralnosti je centralnost stupnjeva (eng. *degree centrality*) koja je jednaka stupnju čvora - drugim riječima, broju veza koje pripadaju danom čvoru. Stupanj čvora može se jednostavno izračunati kao zbroj svih elemenata u određenom retku matrice susjedstva[2]:

$$c_i^{deg} = \sum_j a_{ij}$$

što u matričnom zapisu možemo izraziti kao $C^{deg} = A\mathbf{1}$. $\mathbf{1}$ je $\boldsymbol{N \times 1}$ vektor čiji elementi su jedinice.

Centralnost stupnjeva možemo zamisliti kao poseban slučaj mjere koju je u [4] 1989. predložio Sade, a koja se naziva k-centralnost (eng. *k-path centrality*). Ona broji sve puteve duljine $\boldsymbol{k}$ ili kraće koji izviru iz danog čvora. Prema tome, kada je $\boldsymbol{k = 1}$ mjera postaje identična centralnosti stupnjeva. k-centralnost može se izračunati kao $\boldsymbol{C^{kpath} = W\mathbf{1}}$ gdje je $\boldsymbol{W}$ matrica čiji su elementi $\boldsymbol{w_{ij}}$ broj svih puteva koje duljine $\boldsymbol{k}$ ili kraćih koji spajaju čvorove s indeksom $\boldsymbol{i}$ i $\boldsymbol{j}$.

Varijante k-centralnosti ovise o ograničenjima na tip puteva koje se odluči brojati. Neke od češćih varijanti su:

Geodezijska k-centralnost: Broje se samo geodezijski[3] putevi do duljine $\boldsymbol{k}$ koji izviru iz danog čvora. Na taj način mjerimo utjecaj čvora na geodezijsku strukturu mreže.

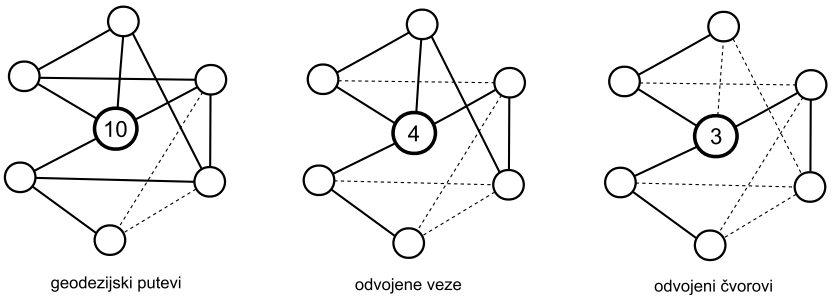
Odvojene veze (eng. edge-disjoint k-path cetrality): Broje se samo putevi koji nemaju zajedničkih veza. Možemo ju također promatrati kao jednu od mjera ranjivosti čvora, jer ona određuje koliko je minimalno čvorova potrebno ukloniti da se prekine veza između dva čvora. Drugi način na koji se ta mjera može promatrati je da je to maksimalni mogući tok[4] između dva čvora.

Odvojeni čvorovi (eng. vertex-disjoint k-path cetrality): Slično kao i u prethodnoj varijanti, s tom razlikom da se ovdje broje samo putevi koji ne sadrže zajedničke čvorove. Njihov broj je jednak minimalnom broju čvorova mreže, koje je potrebno ukloniti da se dva čvora izoliraju jedan od drugoga.

[2] Matrica susjedstva je $N \times N$ kvadratna matrica gdje element a_{ij} ima vrijednost 1 ako između čvorova s indeksima i i j postoji veza, a inače 0. N je broj čvorova u mreži.

[3] *Geodezijski put* je najkraći put između dva čvora.

[4] Ovdje termin *tok* podrazumijeva da svaka veza ima neki kapacitet. U trivijalnom slučaju pretpostavljamo da sve veze u mreži imaju jednak kapacitet.



Slika 1: Varijante 2-centralnosti ovisno o tipu šetnje. Za srednji čvor je naveden broj puteva koji izviru iz njega a poštuju zadano ograničenje. U slučaju 1-centralnosti broj puteva koji izviru iz čvora efektivno postaje stupanj čvora (u ovom slučaju 4).

Freemanova međupoloženost (*Freeman's betweenness centrality*)

Freemanova međupoloženost [5] spada u medijalne mjere centralnosti, što znači da u obzir uzima sve šetnje koje prolaze kroz određeni čvor. Može se reći da je međupoloženost mjera koja nam govori koliko se puta prolazi kroz dani čvor, ako se putuje između dva nasumično odabrana čvora u mreži. Formalno se ona definira s

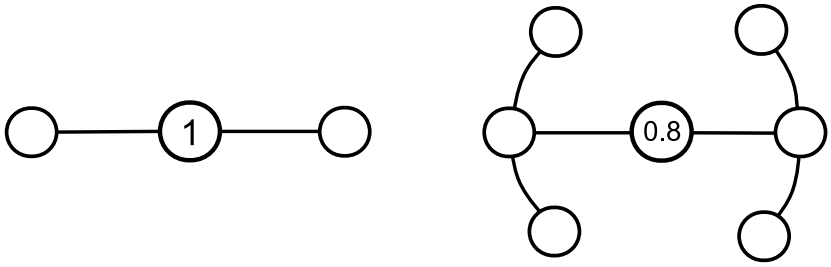
$$C_k^{BET} = \sum_i \sum_j \frac{g_{ikj}}{g_{ij}}$$

gdje je g_{ikj} broj svih geodezijskih puteva koji idu od čvora i do čvora j a prolaze kroz čvor k a g_{ij} broj svih geodezijskih puteva od i do j. Ako u mreži između svaka dva čvora postoji samo jedan najkraći put, međupoloženost je jednaka broju svih najkraćih puteva koji prolaze kroz dani čvor.

Jedna od varijanti Freemanove međupoloženosti je ona u kojoj se u obzir, umjesto samo najkraćih (geodezijskih) puteva, uzimaju svi mogući putevi. Moguće su i ostale varijante koje u obzir uzimaju samo šetnje određenog tipa.

> **k-međupoloženost**: Polazeći od pretpostavke da se putevi velike duljine rijetko koriste, može se ograničiti maksimalna duljina puteva koji se uzimaju u obzir. Tako još 1991. Friedkin u [6] predlaže mjeru koja je efektivno k-međupoloženost s $k = 2$.
>
> **Međupoloženost toka** (eng. flow-betweenness): Brojeći sve puteve (pa i samo one najkraće) između dva čvora u mreži dolazimo u opasnost da višestruko prebrojimo neke veze. Želimo li to izbjeći, možemo odlučiti da prebrajamo samo puteve koji ne dijele zajedničke veze. Već smo rekli da je broj svih puteva koji ne dijele zajedničke veze efektivno količina toka koji se odvija između dva čvora u mreži. Kako flow betweenness mjeri doprinos danog čvora u udjelu svih puteva koji ne dijele zajedničke veze u mreži, može se reći da je on mjera koliko će se smanjiti tok u mreži ako se dani čvor ukloni.

Uglavnom, bez obzira koju varijantu međupoloženosti koristimo, uvjek mjerimo važnost čvora s obzirom na sve moguće šetnje ili tokove u mreži.



Slika 2: Primjer dvije mreže u kojima čvorovi istog stupnja (u ovom slučaju stupnja dva) nemaju jednaku međupoloženost. Riječ je o Freemanovoj međupoloženosti koja je jednaka udjelu svih najkraćih puteva u mreži koji prolaze kroz zadani čvor. Ona ne ovisi o lokalnim svojstvima čvora (kao što je stupanj), već o njegovom položaju unutar strukture mreže.

PageRank algoritam

Jedan od najpoznatijih algoritama za određivanje centralnosti čvorova u mrežama je PageRank kojeg su 1998. na sveučilištu Stanford razvili Lawrence Page i Sergey Brin [3]. Njihova motivacija bila je osmisliti algoritam koji će uspješno rangirati web stranice koje se dobiju kao rezultat upita, a koji se sastoji od jedne ili više ključnih riječi. Dotadašnji su pretraživači jednostavno izlistavali sve stranice u kojima se tražene riječi pojavljuju, ne bivajući u mogućnosti da dane rezultate kvalitetno poredaju prema nekoj mjeri "važnosti" (autoriteta).

Jedan od najjednostavnijih načina da se u kontekstu web stranica ili općenito kod bilo kakve usmjerene mreže, definira važnost stranice, je prebrojiti sve stranice koje se referiraju na danu stranicu. Pošto se web stranice međusobno referiraju pomoću hiperlinkova (eng. *hyperlinks*), te reference se još nazivaju i *ulazni linkovi* (eng. *backlinks*). Intuitivno je jasno da će relevantnije web stranice imati više ulaznih linkova, te se od algoritma očekuje da ih rangira na višu poziciju u odnosu na ostale stranice. Ono što komplicira situaciju je činjenica da samo broj ulaznih linkova nije dovoljan kriterij za procjenu važnoti web stranice. Naime, samo prema tom kriteriju stranica koja ima desetak ulaznih linkova s nerelevantnih stranica (na primjer anonimnih korisničkih foruma) bit će rangirana iznad stranice koja ima samo jedan ulazni link, ali s puno relevantnije i popularnije stranice (na primjer nekog novinskog portala). Znači, u obzir je potrebno uzeti i važnost samih stranica s kojih dolaze ulazni linkovi. Ovo vodi na kružnu definiciju važnosti, koja u biti kaže da je stranica to važnija što su važnije stranice koje se na nju referiraju.

Formalno, problem se svodi na nalaženje svojstvenog vektora (eng. *eigenvector*) $\boldsymbol{R}$ sa svojstvenom vrijednošću $\boldsymbol{c}$ matrice susjedstva $\boldsymbol{A}$ tako da vrijedi $\boldsymbol{R} = \boldsymbol{cAR}$. Implementacija na računalu provodi se tako da se na početku svim stranicama pridjeli neka inicijalna važnost, a onda iterativno proračunava važnosti dok rezultat ne konvergira na neku stabilnu vrijednost i dok gornja jednadžba ne bude zadovoljena.

Kako bi testirali svoj algoritam, Page i Brin razvili su web tražilicu Google. U vrlo kratko vrijeme Google se prometnuo u najpopularniji web pretraživač i uspješnu kompaniju koja je u trenutku izlaska na burzu 2004. godine vrijedila

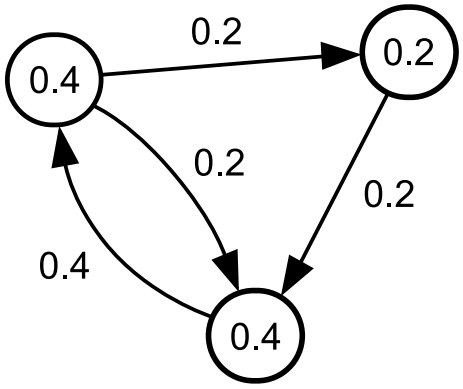
25 milijarde dolara.

Definicija PageRank algoritma

Već je spomenuto da se rang pojedinog čvora u usmjerenoj mreži računa pomoću rekurzivne definicije, koja kaže da je čvor to važniji što su važniji čvorovi koji su vlasnici njegovih ulaznih veza. Formalno se to može iskazati kao

$$\boldsymbol{R(u) = c \sum_{v \epsilon B_u} \frac{R(v)}{N_v}}$$

gdje je $\boldsymbol{u}$ neki čvor, $\boldsymbol{B_u}$ skup njegovih ulaznih veza a $\boldsymbol{N_v}$ broj svih izlaznih veza čvora $\boldsymbol{v}$. Parametar $\boldsymbol{c}$ mora biti manji od $\mathbf{1}$ jer u mreži postoje čvorovi bez izlaznih veza pa se njihov rang gubi iz daljnjeg proračuna (jer ne postoje čvorovi kojima ga raspodijeljuju). Gornja definicija je rekurzivna, no proračun se može provoditi iterativno počevši od bilo kojeg skupa rangova $\boldsymbol{R(u)}$ sve dok vrijednosti ne konvergiraju.



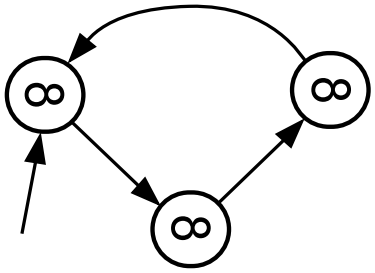
Slika 3: Pojednostavljeni izračun Page Ranka - stacionarno stanje.

Problem sa ovako pojednostavljenom definicijom je što u mreži mogu postojati petlje čvorova koje djeluju kao *ponori ranga* (eng. *rank sink*) pa prikupljaju rang, a nikad ga ne distribuiraju dalje. Zbog toga je potrebno u svakoj iteraciji distribuirati čvorovima određenu količinu dodatnog ranga. Ovo neće utjecati na međusobni poredak čvorova (važniji čvorovi će i dalje imati viši rang u odnosu na ostale), a omogućit će da vrijednosti ostanu normalizirane i da vrijedi $\|\boldsymbol{R}\|_\mathbf{1} = \mathbf{1}$. Uvodimo definiciju *izvora ranga*:

Definicija 1 *Neka je* $\boldsymbol{E(u)}$ *neki vektor čvorova koji korespondira s izvorom ranga. Onda je PageRank skupa čvorova jednakost koja zadovoljava*

$$\boldsymbol{R(u) = c \sum_{v \epsilon B_u} \frac{R(v)}{N_v} + cE(u)}$$

tako da je $\boldsymbol{c}$ *maksimalan i da vrijedi* $\|\boldsymbol{R}\|_\mathbf{1} = \mathbf{1}$ *pri čemu je* $\|\boldsymbol{R}\|_\mathbf{1}$ *1-norma vektora* $\boldsymbol{R}$ [5].



Slika 4: Stranice koje tvore petlju i ne distribuiraju svoj rang dalje - ponor ranga.

U matričnoj notaciji imamo $\boldsymbol{R} = \boldsymbol{c}(\boldsymbol{AR} + \boldsymbol{E})$. Kako vrijedi $\|\boldsymbol{R}\|_1 = \mathbf{1}$ možemo pisati $\boldsymbol{R} = \boldsymbol{c}(\boldsymbol{A} + \boldsymbol{E} \times \mathbf{1})\boldsymbol{R}$ gdje je $\mathbf{1}$ jedinični vektor. Znači, problem nalaženja rangova svih čvorova u mreži $\boldsymbol{R}$ svodi se na nalaženje svojstvenog vektora matrice $(\boldsymbol{A} + \boldsymbol{E} \times \mathbf{1})$.

Vektor $\boldsymbol{E}$ određuje koji čvorovi (i s kolikim udjelom) sudjeluju u obnavljanju ranga koji se distribuira mrežom. U većini slučajeva on može biti uniforman kroz sve čvorove s nekom vrijednosti $\boldsymbol{\alpha}$. $\|\boldsymbol{E}\|_1$ mora biti manji od $\mathbf{1}$ kako bi vrijednosti ranga konvergirale.

Intuitivno, vektor $\boldsymbol{E}$ korespondira s čvorovima kojima je iz nekog razloga dodijeljena veća važnost u odnosu na ostale čvorove. Na primjer, govorimo li o mreži blogova, to bi bili blogovi koje pojedinci iz nekog razloga smatraju relevantnijima i posjećuju učestalije od drugih. Razlozi za takvo ponašanje najčešće su subjektivni - svaki pojedinac ima određeni krug omiljenih blogera čije blogove redovito čita i provjerava. Naravno, ako su njegove vrijednosti uniformne onda nema neke posebne preferencije prema određenim čvorovima.

Ovisno o odabiru vektora $\boldsymbol{E}$ PageRank algoritam će generirati ponekad i vrlo različite rangove. Pri tome će čvorovi čije komponente u vektoru imaju veću vrijednost biti preferirani, s odgovarajuće višim PageRankom od ostalih. Ta karakteristika se može iskoristiti ako se primjerice želi definirati "personalizirana" centralnost u kojoj će određena skupina čvorova i njihove veze biti preferirani. Jedna od primjena je pri pretraživanju web stranica gdje se pretraživaču želi sugerirati koje su stranice, prema korisnikovom mišljenju, povezane s traženim upitom.

Izračun PageRanka

Izračunavanje vrijednosti ranga za sve čvorove provodi se iterativno sljedećim algoritmom:

[5] Ako komponente vektora $\boldsymbol{R}$ označimo s $\boldsymbol{k_1, k_2, k_3, \ldots, k_n}$ onda je i njegova 1-norma jednaka $|\boldsymbol{k_1}| + |\boldsymbol{k_2}| + |\boldsymbol{k_3}| + \cdots + |\boldsymbol{k_n}|$

$\boldsymbol{R_0} \leftarrow \boldsymbol{S}$
`ponavljaj:`

$\boldsymbol{R_{i+1}} \leftarrow \boldsymbol{AR_i}$
$\boldsymbol{d} \leftarrow \|\boldsymbol{R_i}\|_1 - \|\boldsymbol{R_{i+1}}\|_1$
$\boldsymbol{R_{i+1}} \leftarrow \boldsymbol{R_{i+1}} + \boldsymbol{dE}$
$\boldsymbol{\delta} \leftarrow \|\boldsymbol{R_{i+1}} - \boldsymbol{R_i}\|_1$

`dok` $\boldsymbol{\delta} > \boldsymbol{\epsilon}$

pri čemu je $\boldsymbol{S}$ vektor početnih vrijednosti rangova za sve čvorove, a $\boldsymbol{d}$ faktor koji povećava brzinu konvergencije i održava $\|\boldsymbol{R}\|_1 = \mathbf{1}$. Matrica susjedstva $\boldsymbol{A}$ definirana je kao

$$\boldsymbol{A} = \begin{cases} \frac{\mathbf{1}}{\boldsymbol{N_u}}, & \text{ako postoji veza} \\ \mathbf{0}, & \text{inače} \end{cases}$$

gdje je $\boldsymbol{N_u}$ broj izlaznih veza čvora $\boldsymbol{u}$.

2 Mjera centralnosti temeljena na odzivu

U određivanju ključnih čvorova od presudne je važnosti poznavati strukturu mreže tj. odnose njenih veza i čvorova. Jednom kad je ta struktura poznata ključni čvorovi se mogu odrediti raznim metodama, ovisno već o definiciji centralnosti. Za procjenu centralnosti pojedinog čvora potrebno je poznavati strukturu cijele mreže, kao što je to slučaj kod Freemanove međupoloženosti i PageRanka, ili samo dio, u k-centralnosti na primjer.

Opravdano se postavlja pitanje mogu li se na neki način ključni čvorovi odrediti bez poznavanja strukture mreže? Naravno, ako su dostupne samo informacije o čvorovima i vezama (dakle, o strukturi mreže), onda je odgovor negativan. No ne postoji razlog zašto bi se ključna svojstva mreže, pa tako i centralnost, pokušavala predvidjeti samo iz njene strukture. I druge informacije, nevezane za strukturu mreže, također se mogu iskoristiti za identificiranje ključnih čvorova istovremeno vodeći računa da principi budu dovoljno općeniti kako bi se mogli primjeniti i na mreže iz drugih domena.

Predložene su dvije mjere koje svoj pristup ne temelje direktno na lokalnoj ili globalnoj strukturi mreže. Ovdje će one biti opisane u kontekstu mreže blogova, no vjerojatno adekvatne mjere postoje i za bilo kakvu mrežu u kojoj je moguće definirati svojstvo odziva.

> **Unutarnji signal vrha**: Odgovara broju, duljini i ostalim karakteristikama svakog posta na pojedinom blogu. Orijentiran je isključivo na informaciju koja je vezana za samog blogera (na koji način i kako piše svoje tekstove) pa upitan ostaje odnos na ostale čvorove u mreži. U okviru ovog istraživanja unutarnji signal vrha nije dodatno razmatran.
>
> **Vanjski signal vrha**: Odgovara broju, duljini i ostalim karakteristikama svih komentara na pojedine postove na blogu. Kako je orijentiran na komentare drugih blogera u njemu se indirektno oslikava utjecaj pojedinog blogera na ostale blogere tj. ostatak mreže. Vanjski signal vrha, u nastavku rada skraćeno će se nazivati samo *odziv*. Takav naziv sugerira povratnu reakciju koja se događa kad jedan bloger komentira post drugog blogera. Ova reakcija ima važnu ulogu u definiranju autoriteta koji se pridaje pojedinim blogerima. Naime, u izostanku objektivnih kriterija koji bi vrednovali kvalitetu samih postova, informacija o broju, duljini i vrsti komentara ispostavlja se kao relevantna mjera o subjektivnoj važnosti pojedinog blogera. Blogere čiji tekstovi nisu čitani, sasvim sigurno ne bi nazvali popularnima - terminom kojim se u socijalnim mrežama često referira na centralnost unutar strukture mreže. Pretpostavka je da će mjere temeljene na strukturalnom pristupu blogere s velikim odzivom, također visoko rangirati.

Opravdanje za ovakav pristup je da, koliko god pojam centralnosti bio vezan za strukturu same mreže, faktori koji dovode do toga da neki čvor bude "centralniji" od drugog to nisu. Svaka realna mreža mora proći kroz proces evolucije u kojem se stvara njezina struktura i koji njoj i njenim čvorovima daje određena svojstva i ulogu u mreži - radi li se o slabo ili jako povezanom čvoru, povezuje li on dijelove koji bi inače bili razdvojeni i tako dalje... Barabási i Albert u

[7] pokazuju kako preferencijalno povezivanje[6] u rastućim mrežama dovodi do svojstava kao što je distribucija zakona potencija (eng. *power-law degree distribution*).

Modeli mjere odziva

Mjere temeljene na odzivu jedan su od pokušaja da se važnost čvora definira lokalnim parametrima samoga čvora koje nisu direktno vezane za strukturu. U slučaju mreže blogova važne lokalne karakteristike su:

- Broj i duljina svakog posta
- Broj i duljina svakog komentara

Notacija koja će se koristiti je:

$\boldsymbol{N}$ - ukupni broj svih blogova u mreži
$\boldsymbol{R(u)}$ - odziv bloga sa indeksom $\boldsymbol{u}$
$\boldsymbol{C(u)}$ - skup duljina svih komentara na blog sa indeksom $\boldsymbol{u}$
$\boldsymbol{C_i(u)}$ - duljina pojedinog komentara na blog sa indeksom $\boldsymbol{u}$
$\boldsymbol{|C(u)|}$ - broj svih komentara na blog sa indeksom $\boldsymbol{u}$
$\boldsymbol{|C(u,p)|}$ - broj svih komentara na post s indeksom $\boldsymbol{p}$ koji pripada blogu s indeksom $\boldsymbol{u}$
$\boldsymbol{P(u)}$ - skup duljina svih postova bloga sa indeksom $\boldsymbol{u}$
$\boldsymbol{P_i(u)}$ - duljina pojedinog posta bloga sa indeksom $\boldsymbol{i}$
$\boldsymbol{|P(u)|}$ - broj svih postova bloga sa indeksom $\boldsymbol{u}$

Definiramo devet mjera temeljenih na odzivu $\boldsymbol{R(u)}$:

I. model

$$\boldsymbol{R(u) = \frac{\sum_{i=1}^{N} C_i(u)}{\sum_{i=1}^{N} P_i(u)}}$$

II. model

$$\boldsymbol{R(u) = \frac{\sum_{i=1}^{N} C_i(u)}{|P(u)|}}$$

III. model

$$\boldsymbol{R(u) = \frac{|C(u)|}{|P(u)|}}$$

IV. model

$$\boldsymbol{R(u) = \frac{\sum_{i=1}^{N} C_i^2(u)}{\sum_{i=1}^{N} P_i(u)}}$$

V. model

$$\boldsymbol{R(u) = \frac{\sum_{i=1}^{N} C_i^2(u)}{|P(u)|}}$$

[6] Preferencijalno povezivanje (eng. *preferential attachment*) znači da čvorovi velikog stupnja imaju veću vjerojatnost privlačenja novih čvorova i time priliku za dodatno povećanje svojeg stupnja. Taj mehanizam omogućuje stvaranje određenog broja čvorova vrlo velikog stupnja, takozvanih habova (eng. *hubs*). U kolokvijalnom izričaju za preferencijalno povezivanje se koriste i nazivi "bogati postaju bogatiji", "pobjednik dobiva sve"...

VI. model

$$R(u) = \frac{\sum_{i=1}^{N} C_i^2(u)}{\sum_{i=1}^{N} P_i(u)} + \sum_{i=1}^{N} C_i(u)$$

VII. model

$$R(u) = \sum_{i=1}^{N} C_i^2(u)$$

VIII. model

$$R(u) = \frac{|C(u)|^2}{|P(u)|} + |C(u)|$$

IX. model

$$R(u) = \frac{\sum_{p=1}^{|P(u)|} |C(u,p)|^2}{|P(u)|}$$

3 Kompleksna mreža blogova

Za potrebe analize prethodno izloženih algoritama i metoda modelirana je kompleksna mreža temeljena na blogovima. Koncetriralo se na jedan blog servis - `www.bloger.hr` koji se pokazao iznimno pogodnim zbog velikog broja blogova (preko 13000) ali i vrlo urednog HTML koda iz kojeg je bilo jednostavno dohvatiti potrebne podatke. Pri tome se blogeri (korisnici koji pišu blogove) gledaju kao čvorovi mreže, a njihovi međusobni komentari kao veze između njih.

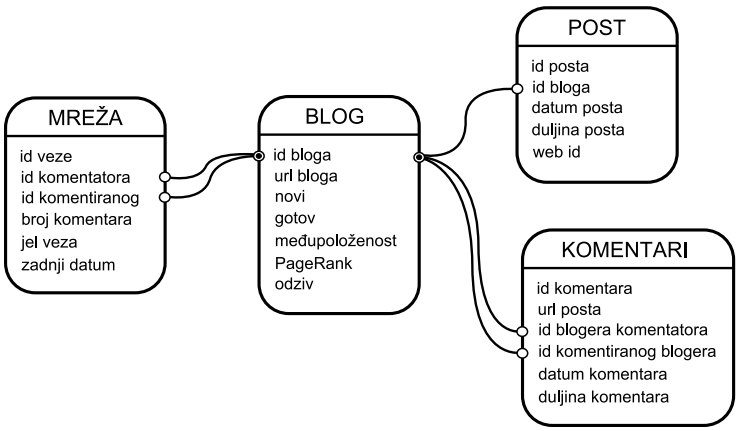
Dohvaćanje podataka

Za dohvaćanje podataka koristio se programski jezik Perl (Practical Extraction and Report Language) i MySql baza podataka. Riječ je o otvorenim tehnologijama besplatno dostupnim na webu s lako dostupnom i opširnom dokumentacijom. Perl se pokazao kao iznimno pogodan za izgradnju tekstualnog parsera koji je uspješno iz gomile HTML koda izvukao bitne podatke i pospremio ih u bazu. Informacije koje su bile potrebne za izgradnju mreže su:

Blogovi: url adresa

Postovi: duljina u znakovima, datum

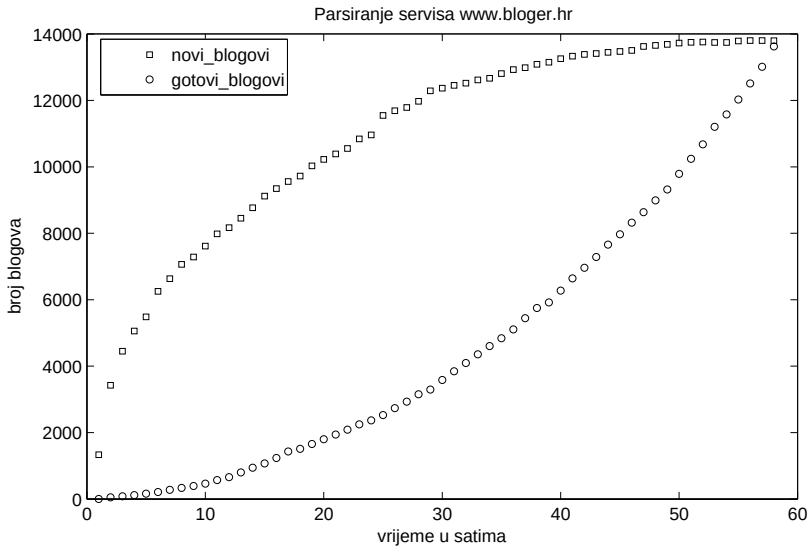
Komentari na postove: duljina u znakovima, vlasnik komentara, kome je komentar upućen, datum



Slika 5: Model baze podataka u koju su pohranjeni podatci o mreži blogova. Linijama su prikazane reference (strani ključevi), kojima su povezani atributi u različitim tablicama.

Postupak parsiranja podataka započet je 19. svibnja 2008. i trajao je nešto manje od 70 sati. Dohvaćali su se svi podatci o blogovima, postovima i komentarima od siječnja 2004. godine (od kad postoji `bloger.hr`). Tijek parsiranja može se vidjeti na slici 6. Parsiranje počinje od nekog početnog bloga koji se ručno pohrani u bazu. Potom se dohvate svi njegovi postovi i imena blogera koji su ih komentirali. Adrese novih blogova[7] se slijedno pohranjuju u bazu

[7] Pošto se u obzir uzimaju samo blogeri s istog blog servisa svaki bloger i adresa njegovog bloga je jednoznačno određena imenom samog blogera.



Slika 6: Napredak parsiranja kroz vrijeme - novi blogovi i gotovi blogovi.

gdje će ih parser naknadno obraditi. Postupak se ponavlja dokle god ima novih blogova.

U početku se nalazi mnogo novih blogova koji redovito pripadaju aktivnim blogerima s puno postova i komentara pa samo parsiranje teče sporije. Kako parsiranje napreduje, dohvaćaju se manje popularni blogovi s manje postova i komentara pa je potrebno kraće vrijeme da ih se parsira. Istovremeno, većina registriranih blogova je već nađena pa se dodavanje novih blogova u bazu usporava. To nam je ujedno i garancija da će postupak parsiranja završiti u konačnom vremenu (uostalom, i broj registriranih blogova na servisu je konačan).

Samo dohvaćanje podataka implementirano je iterativno - to znači da se ne dohvaćaju podatci za one blogove, postove i komentare koji se već nalaze u bazi. Zato se parsiranje cijelog servisa mora provesti samo jednom, a svako sljedeće dohvaćat će samo podatke koji su se pojavili u međuvremenu.

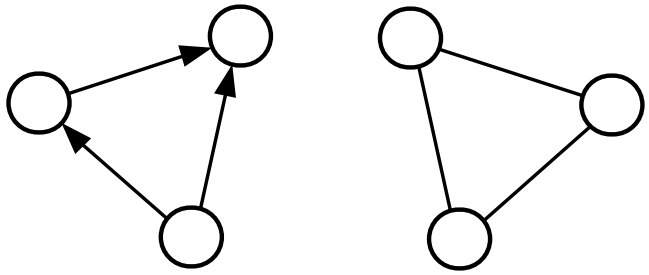
Izgradnja kompleksne mreže blogova

Jednom kad su dohvaćeni svi potrebni podatci može se krenuti na izgradnju dviju mreža - *usmjerene* i *neusmjerene*.

> **Usmjerena mreža**: Svaka veza između dva blogera ima definirani smjer. Tako je za svakog blogera moguće razlikovati izlazne (eng. *forward links*) i ulazne veze (eng. *backlinks*). Bloger dobiva izlaznu vezu u trenutku kada komentira nekog drugog blogera, a ulaznu vezu u trenutku kad drugi bloger komentira njega.
>
> **Neusmjerena mreža**: Veze nemaju definirani smjer i postoji samo informacija postoji li između dva blogera veza ili ne. Za razliku od

usmjerene veze, koja je u mreži blogova sama po sebi jasna, *neusmjerenu vezu* je potrebno dodatno definirati. Za potrebe ovog rada odabran je model u kojem neusmjerena veza između dva blogera postoji ako su obojica barem jednom komentirali jedan drugoga. Takav način definiranja veze garantira da je ona (makar indirektno) stvorena pristankom oba blogera.

Usmjerena mreža važna je za mjere koje u obzir uzimaju razliku između izlaznih i ulaznih veza, kao što je na primjer PageRank. Većina ostalih, uključujući k-centralnost, Freemanovu međupoloženost i odziv, oslanja se na neusmjerenu mrežu gdje ta distinkcija nije bitna.



Slika 7: Primjer jednostavne usmjerene (lijevo) i neusmjerene mreže (desno).

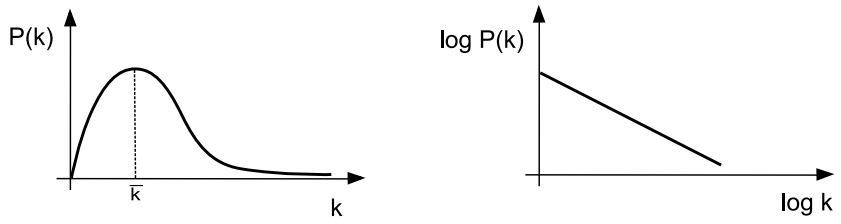
Parametri dobivene kompleksne mreže

Za dobivenu mrežu računaju se distribucije stupnjeva za ulazne i izlazne veze, te distribucija za neusmjerenu mrežu i kumulativna distribucija. Stupanj čvora je zapravo broj veza koji posjeduje. Naravno, način definiranja veze ovisi o kontekstu - ako je riječ o neusmjerenoj mreži, stupanj čvora zapravo je ništa drugo do poseban slučaj Sadeove k-centralnosti u kojoj je $\boldsymbol{k} = \mathbf{1}$. U usmjerenim mrežama, gdje se razlikuju izlazne i ulazne veze, potrebno je definirati i dvije distribucije stupnjeva - jednu za izlazne i jednu za ulazne veze.

Distribucija stupnjeva je vjerojatnost da će neki čvor u mreži veličine $\boldsymbol{N}$ čvorova imati stupanj $\boldsymbol{k}$. Možemo govoriti i o *kumulativnoj distribuciji* - u tom slučaju nas zanima vjerojatnost da će neki čvor imati stupanj *veći* od $\boldsymbol{k}$. Intuitivno očekujemo da će u mreži blogova većina blogera u prosjeku imati neki srednji stupanj, a da će vjerojatnost nalaženja blogera sa stupnjem koji znatno odskače od toga (bilo da je premali ili preveliki) biti vrlo mala. Takvu situaciju dobro opisuje Poissonova distribucija koja ima maksimum kod neke srednje vrijednosti stupnja $\overline{\boldsymbol{k}}$, a za ostale vrijednosti pada.

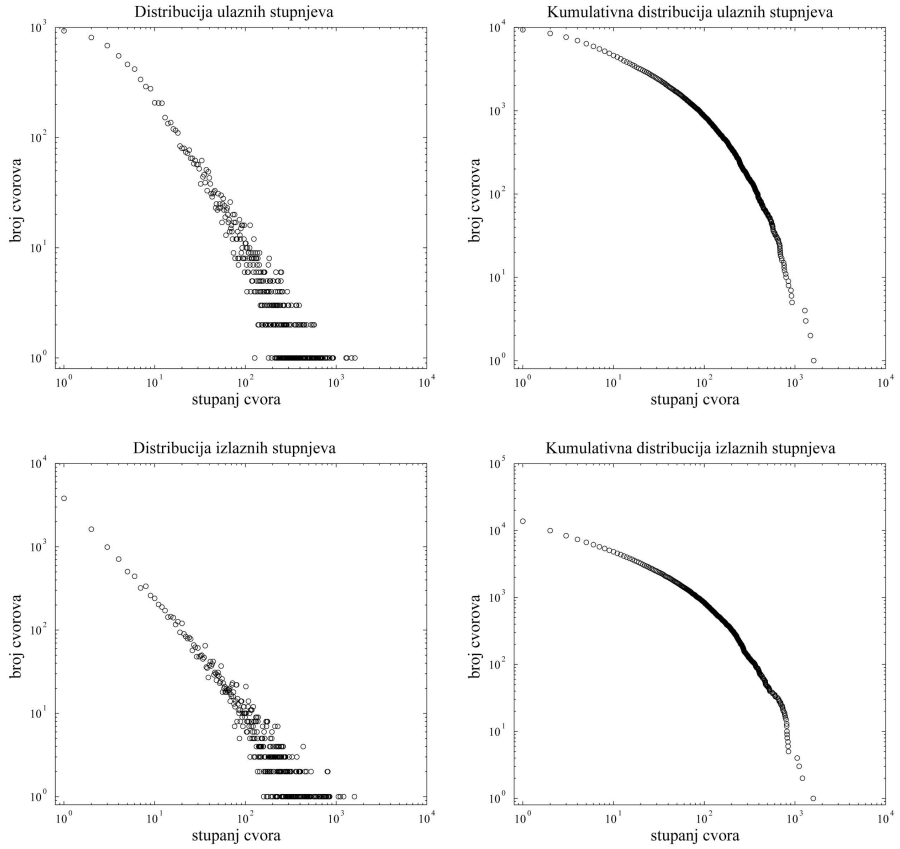
Ipak, nakon što je napretkom informacijske tehnologije postalo moguće analizirati mreže s vrlo velikim brojem čvorova (kao što su na primjer Internet, World Wide Web ili mreže proteinskih interakcija) pokazalo se da većina stvarnih mreža ne prati takvu Poissonovu distribuciju stupnjeva. U stvarnim mrežama, uz veliki broj čvorova niskog stupnja, također postoje i čvorovi čiji stupanj za par redova veličine premašuje prosječni stupanj mreže $\overline{\boldsymbol{k}}$. Takve mreže dobro opisuje distribucija zakona potencija (eng. *power law distribution*). Barabasi i Albert

su 1999. u [7] prvi uspjeli razviti model koji uspješno opisuje nastanak mreža s distribucijom stupnjeva koja odgovara zakonu potencija.



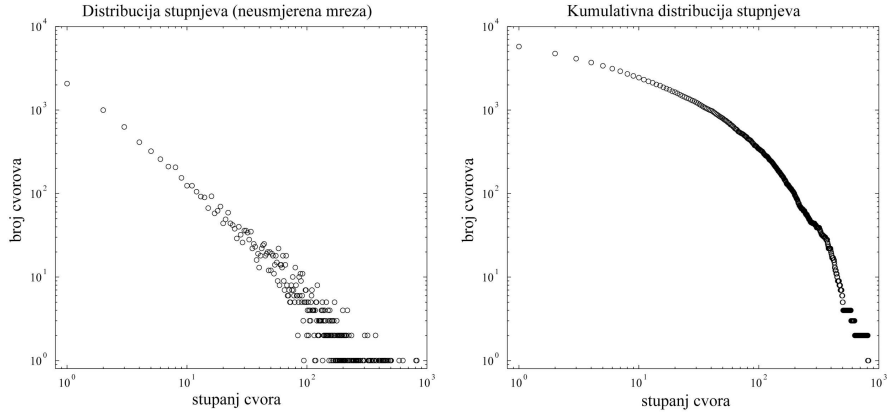
Slika 8: Poissonova distribucija i distribucija zakona potencija.

Alternativno se umjesto vjerojatnosti pojavljivanja čvora nekog stupnja može prikazati broj čvorova tog stupnja. To mijenja mjerne jedinice na osi y no ne i izgled distribucije[8].



Slika 9: Distribucije za ulazne i izlazne stupnjeve mreže bloger.hr.

[8]Iako to onda više nije distribucija u formalnom smislu.



Slika 10: Distribucija stupnjeva za neusmjerenu mrežu bloger.hr.

4 Identifikacija ključnih igrača u kompleksnoj mreži blogova

Sada se na dobivenim modelima mreža mogu provesti metode za traženje ključnih igrača. Od opisanih strukturalnih mjera implementirano je rangiranje prema stupnju čvora i PageRank. Odabrane su upravo te dvije mjere, jer one pristupaju problemu centralnosti iz *lokalne* i *globalne* perspektive.

> **Lokalna struktura**: Procjena centralnosti radi se na temelju lokalne strukture za koju nije potrebno poznavati strukturu cijele mreže. Ovdje se stupanj čvora činio kao pogodna početna mjera.
> **Globalna struktura**: Procjena centralnosti radi se na temelju strukture cijele mreže - dakle, globalno. Kao pogodna mjera odabran je PageRank, prije svega zbog jednostavnije implementacije i bržeg izračuna za mrežu s velikim brojem čvorova u usporedbi s međuloženosti toka.

Također, implemenirano je i svih devet modela odziva.

Izračun stupnja čvora za blogove

U bazi podataka čuvaju se podatci o tome koji je bloger komentirao kojeg blogera. Već samo ti podatci mogu se iskoristiti za stvaranje neusmjerene mreže u kojoj se komentar jednog blogera na drugog broji kao jedna neusmjerena veza. Ipak, u ovom modelu su zbog jednostavnosti višestruke veze grupirane i naveden je samo njihov broj. Treba obratiti pozornost na to da su takve veze i dalje usmjerene.

Kako će se analiza stupnjeva čvorova provoditi nad neusmjerenom mrežom, potrebno je pregledati sve usmjerene veze i odabrati one koje zadovoljavaju uvjete da postanu neusmjerene. U našem slučaju je uvjet da postoji barem jedna usmjerena veza od jednog do drugog blogera i jedna natrag od drugog do prvog. Tada će za svaku vezu biti dostupan podatak koja dva blogera ju tvore ali više neće biti važno u kojem je smjeru orijentirana. Stupanj čvora se potom lako izračuna i to jednostavnim prebrojavanjem koliko se puta neki čvor pojavio u tablici neusmjerenih veza.

Najbolje rangiranih 10 blogova prema broju veza su:

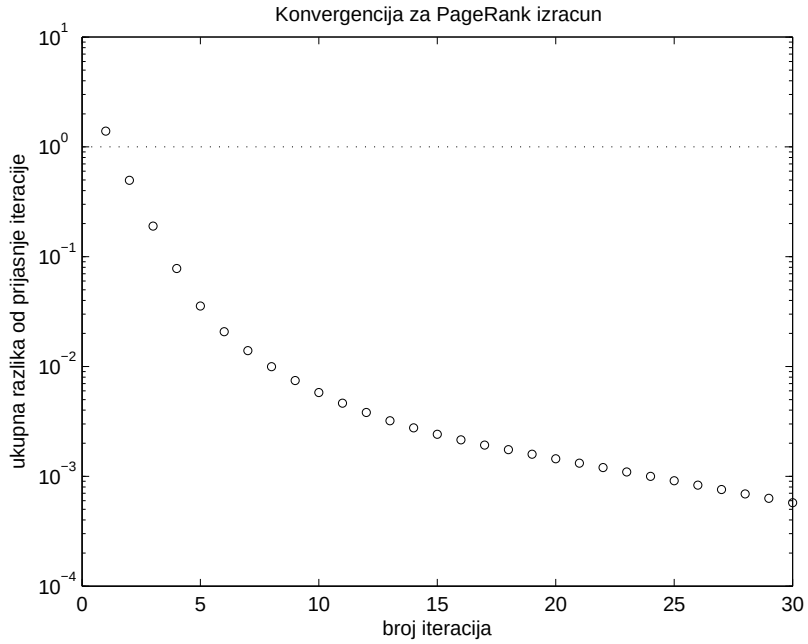
BLOG	STUPANJ
1. http://vigga.bloger.hr	821
2. http://pjegavi.bloger.hr	811
3. http://meg.bloger.hr	629
4. http://moje.bloger.hr	586
5. http://ludonjabravo.bloger.hr	503
6. http://tajnatajna.bloger.hr	498
7. http://jjbraddock.bloger.hr	492
8. http://snjezana_leona.bloger.hr	478
9. http://vedriosmjeh.bloger.hr	468
10. http://auroraisaa.bloger.hr	463

Izračun PageRanka za blogove

Algoritam koji se koristio za izračun PageRanka nalazi se u poglavlju 1. On se provodi na usmjerenoj mreži. Vrijednosti vektora $\boldsymbol{E}$ su jednake za sve čvorove i vrijedi $\|\boldsymbol{E}\|_1 = \mathbf{1}$. Proračun se provodi u 30 iteracija što garantira da će greška postati dovoljno mala da više ne utječe znatno na iznose PageRanka. Apsolutnu razliku[9] između trenutnog ranga i ranga u prijašnjoj iteraciji može se vidjeti na slici 11.

Deset najbolje rangiranih blogova prema PageRanku su:

BLOG	PAGERANK
1. http://naslovnica.bloger.hr	0.0100858406693737
2. http://matija.bloger.hr	0.0045955180832378
3. http://vigga.bloger.hr	0.00420363200298373
4. http://lana.bloger.hr	0.00405655764802687
5. http://pjegavi.bloger.hr	0.00387382699172099
6. http://Pearl.bloger.hr	0.00322977024527641
7. http://NeMresBilivit.bloger.hr	0.00306041019415207
8. http://vin.bloger.hr	0.0030318123447316
9. http://leonna.bloger.hr	0.00300944690409522
10. http://marinshe.bloger.hr	0.00278712597571489



Slika 11: Razlika između PageRank vrijednosti za svaki čvor smanjuje se u svakoj iteraciji što ukazuje da se proračun približava stacionarnom stanju.

[9] Misli se, naravno, na razliku $\|\boldsymbol{R_{i+1}} - \boldsymbol{R_i}\|_1$ koja izražava promjenu PageRank vrijednosti u susjednim iteracijama.

Izračun odziva za blogove

U poglavlju 2 opisano je devet modela mjere odziva. Za svaku od tih mjera provedeni su zasebni izračuni. Podatci o broju i duljini svih komentara i postova dohvaćani su iz baze podataka. Vrijednosti tako dobivenih odziva normirane su pomoću

$$\boldsymbol{R'(u)} = \frac{\boldsymbol{R(u)}}{\|\boldsymbol{R}\|_{\mathbf{1}}}$$

gdje je $\boldsymbol{R(u)}$ vrijednost odziva za blog s indeksom $\boldsymbol{u}$ a $\|\boldsymbol{R}\|_{\mathbf{1}}$ 1-norma vektora $\boldsymbol{R}$.

5 Usporedba dobivenih rezultata

Jednostavno izlistavanje $\boldsymbol{n}$ najboljih rezultata za neku mjeru centralnosti, ne govori nam puno o tome koliko su rezultati vjerodostojni. Primjerice, ravnamo li se prema rezultatima sa servisa `www.bloger.hr`, najpopularnijih 10 blogova na dan 5. lipnja 2008. su:

1. `http://lana.bloger.hr`
2. `http://suton_.bloger.hr`
3. `http://tose.bloger.hr`
4. `http://kawai.bloger.hr`
5. `http://sasava.bloger.hr`
6. `http://stereosound.bloger.hr`
7. `http://acm1989.bloger.hr`
8. `http://folkshowbiz.bloger.hr`
9. `http://NeMresBilivit.bloger.hr`
10. `http://sexangelina.bloger.hr`

Najvažniji kriterij prema kojem se blogovi rangiraju je njihova posjećenost. Ako bloger tjedan dana ne objavi nijedan post, njegov blog se skida s liste. Ovakav način rangiranja preferira trenutno aktivne blogere i dobro funkcionira u dinamičkom okruženju, kao što je zajednica blogera. Podatci o posjećenosti pojedinih blogova, nisu bili dostupni za ovo istraživanje, iako bi bilo zanimljivo provjeriti kako oni koreliraju s mjerom odziva i ostalim mjerama centralnosti temeljenim na strukturalnom pristupu. U svakom slučaju, vremenski raspon od samo tjedan dana u kojem se mjeri posjećenost, sasvim sigurno nije pouzdan pokazatelj stvarne važnosti čvora, ako nam je u interesu centralnost definirati kao svojstvo koje se neće znatno promijeniti u situaciji kada pojedini bloger na tjedan dana prestane pisati postove.

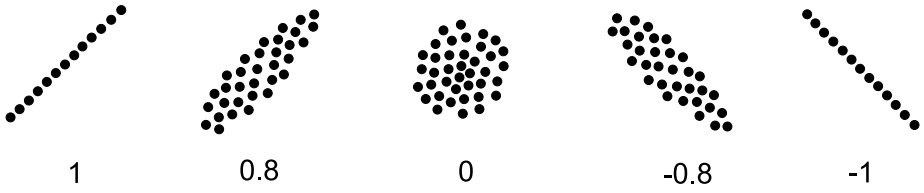
Provedene su dvije vrste usporedbi - usporedba *rangova* i usporedba *vrijednosti*.

> **Usporedba rangova**: U xy ravnini plotane su sve vrijednosti rangova prema ispitivanim mjerama odziva. Na primjer, ako je određeni bloger rangiran na 10. poziciju prema jednoj mjeri i 25. prema drugoj onda njega predstavlja točka s koordinatom (10,25). Za sve usporedbe naveden je koeficijent korelacije $\boldsymbol{r}$ koji sažeto izražava kvalitetu korelacije. Na slici 12 vidimo da bi u slučaju idealne korelacije ($\boldsymbol{C = 1}$) sve točke na grafu bi bile smještene na pravcu $\boldsymbol{y = x}$.
> **Usporedba vrijednosti**: Slično kao i kod usporedbe rangova i ovdje se vrijednosti plotaju u xy ravnini. Zbog velikog raspona vrijednosti korištena je logaritamska skala. Kako same vrijednosti dviju mjera najčešće nisu usporedne, ne izračunava se koeficijent korelacije.

Grafikoni su generirani uz pomoć programskog paketa Matlab. Za izračun koeficijenta korelacije $\boldsymbol{r}$ korištena je Matlab naredba `corr` koja u standardnom načinu rada koristi Pearsonov koeficijent korelacije

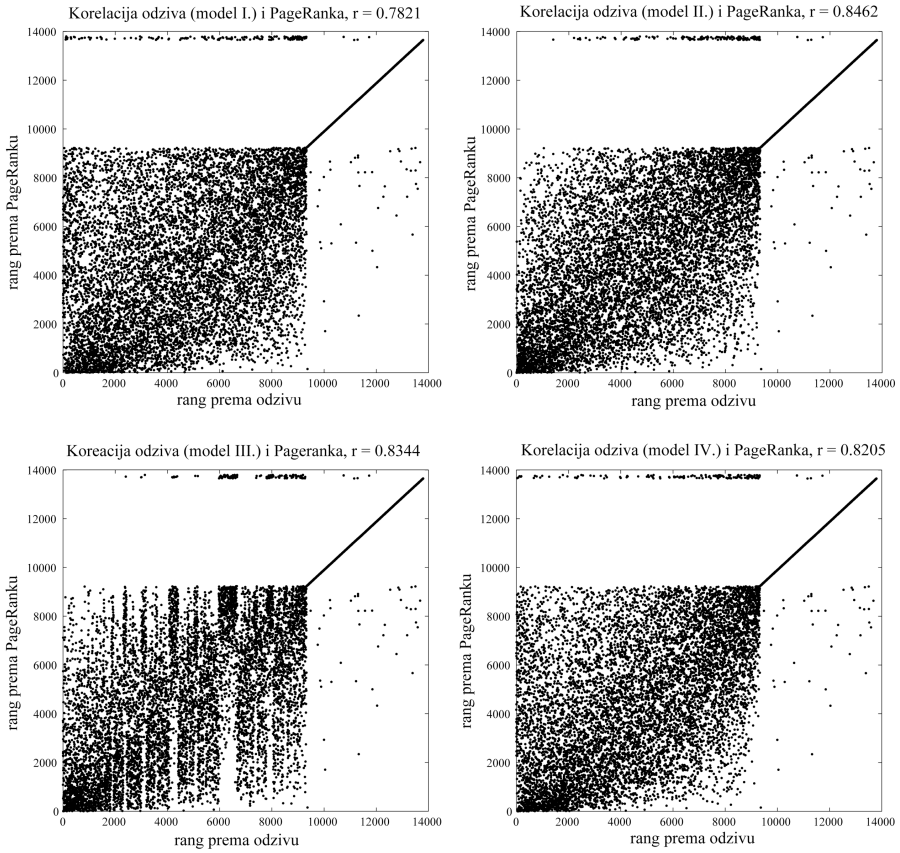
$$\boldsymbol{r_{X,Y}} = \frac{\mathrm{cov(X,Y)}}{\boldsymbol{\sigma_X \sigma_Y}}$$

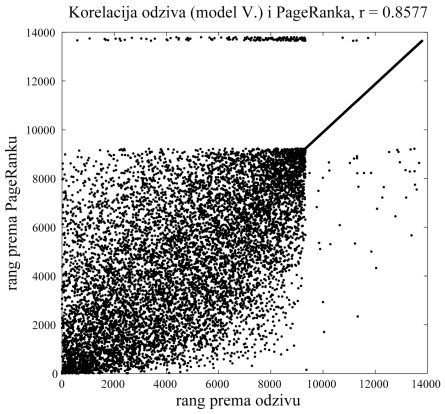
gdje je cov(X,Y) kovarijacijski moment, a $\boldsymbol{\sigma_X}$ i $\boldsymbol{\sigma_Y}$ standardne devijacije dviju slučajnih varijabli $\boldsymbol{X}$ i $\boldsymbol{Y}$.

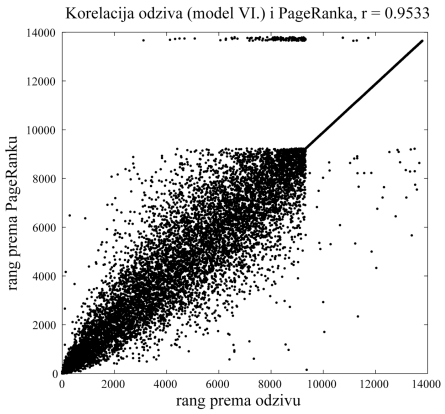


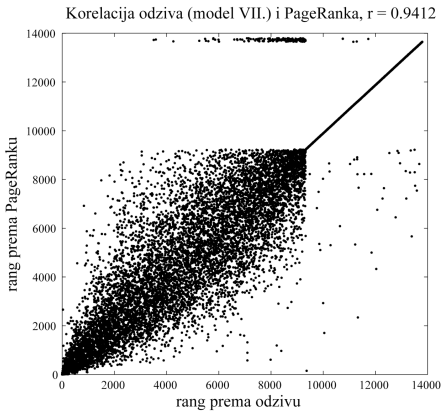
Slika 12: Primjer skupova (x,y) vrijednosti s okvirnim koeficijentima korelacije. Koeficijent korelacije ovisi o orijentaciji (pozitivna ili negativna) i raspršenosti u linearnom smjeru

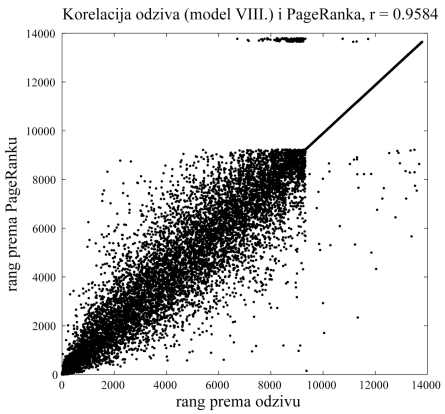
Usporedba rangova dobivenih odzivom i PageRankom

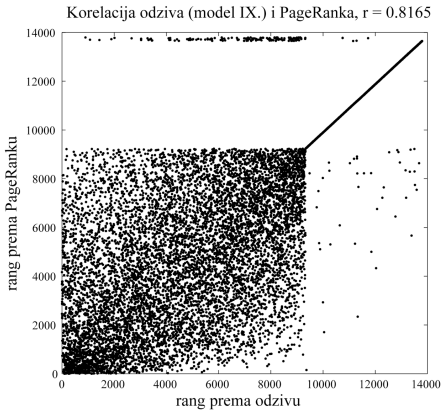



Korelacija odziva (model V.) i PageRanka, r = 0.8577
rang prema PageRanku
rang prema odzivu
0
2000
4000
6000
8000
10000
12000
14000

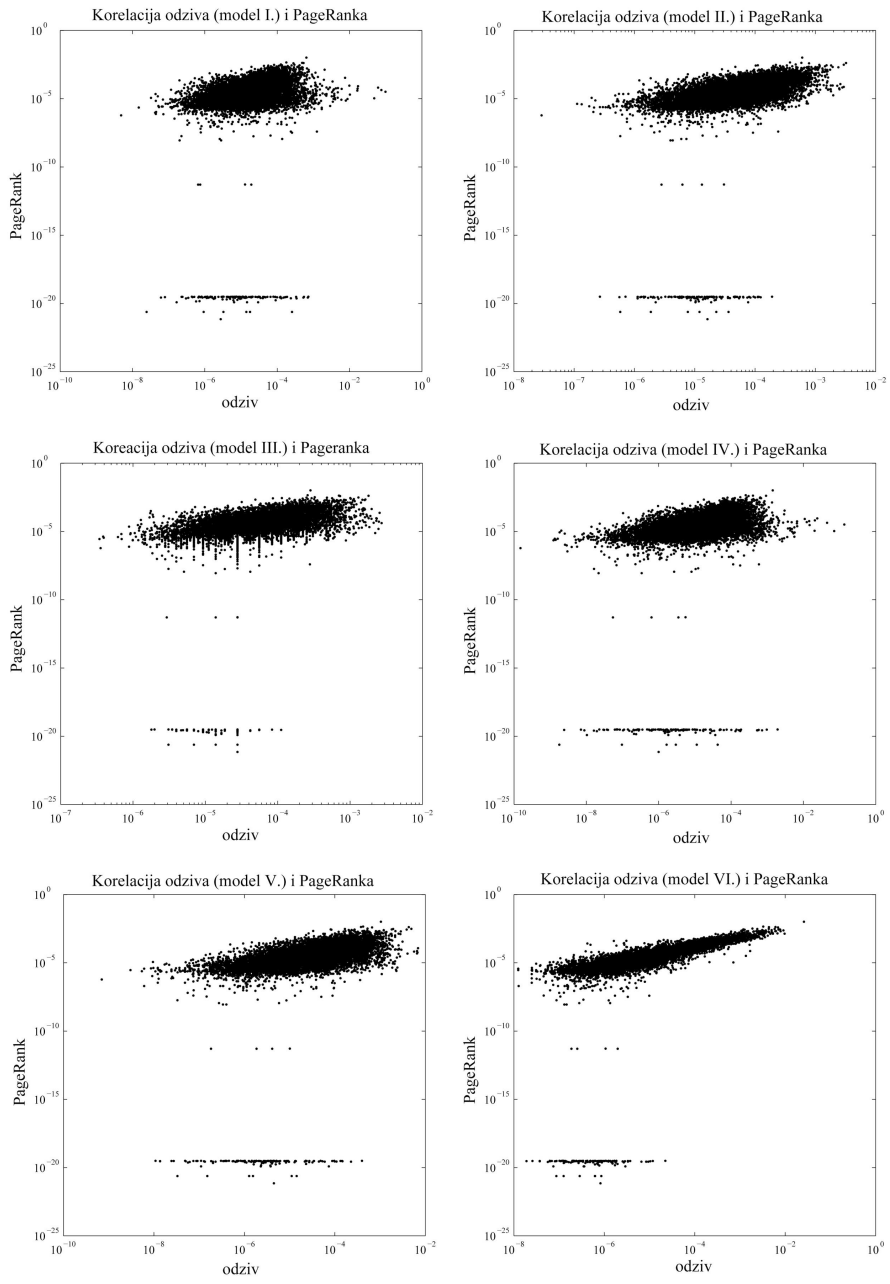

Korelacija odziva (model VI.) i PageRanka, r = 0.9533
rang prema PageRanku
rang prema odzivu
0
2000
4000
6000
8000
10000
12000
14000

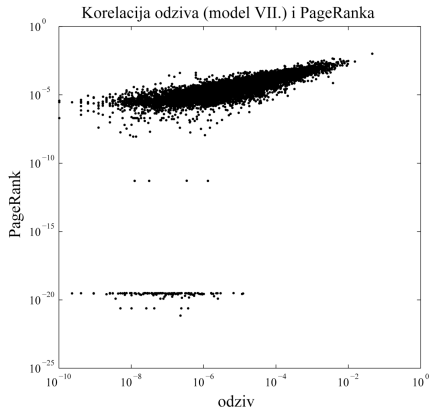

Korelacija odziva (model VII.) i PageRanka, r = 0.9412
rang prema PageRanku
rang prema odzivu
0
2000
4000
6000
8000
10000
12000
14000

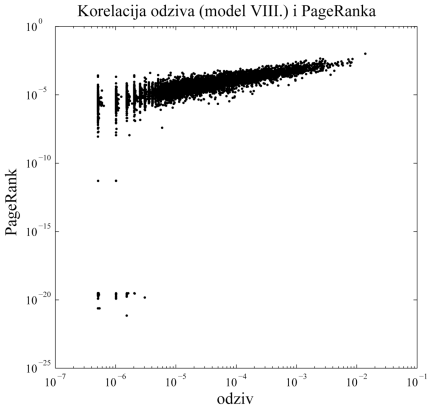

Korelacija odziva (model VIII.) i PageRanka, r = 0.9584
rang prema PageRanku
rang prema odzivu
0
2000
4000
6000
8000
10000
12000
14000

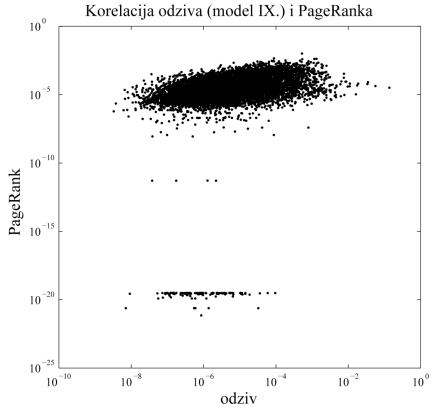

Korelacija odziva (model IX.) i PageRanka, r = 0.8165
rang prema PageRanku
rang prema odzivu
0
2000
4000
6000
8000
10000
12000
14000

Usporedba vrijednosti dobivenih odzivom i PageRankom

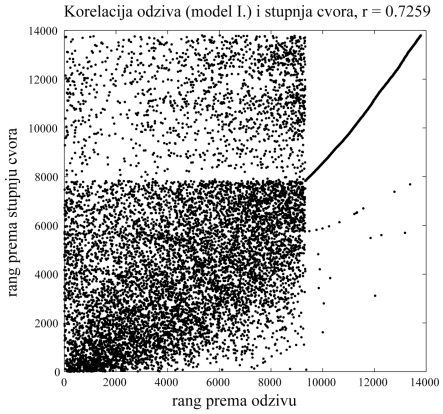


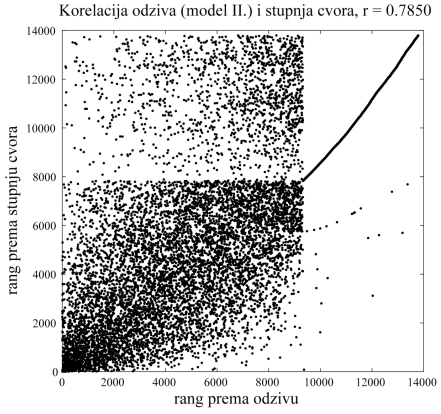

Korelacija odziva (model VII.) i PageRanka
PageRank
odziv

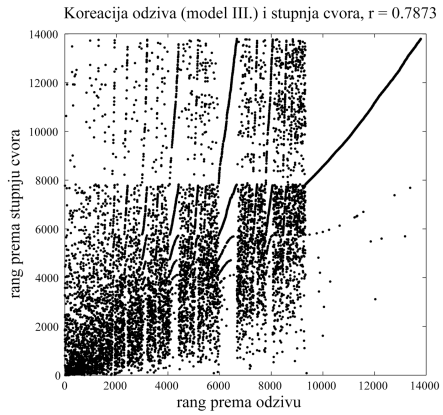

Korelacija odziva (model VIII.) i PageRanka
PageRank
odziv

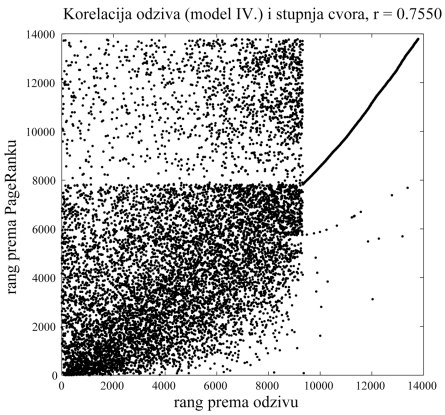

Korelacija odziva (model IX.) i PageRanka
PageRank
odziv

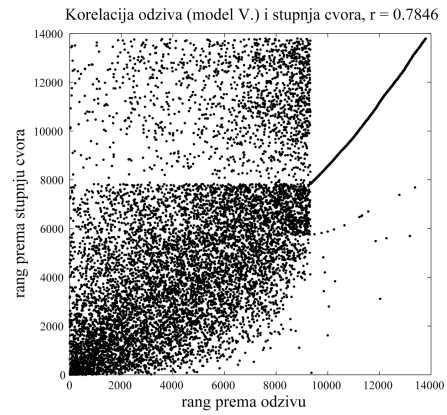
Usporedba rangova dobivenih odzivom i stupnjem čvora

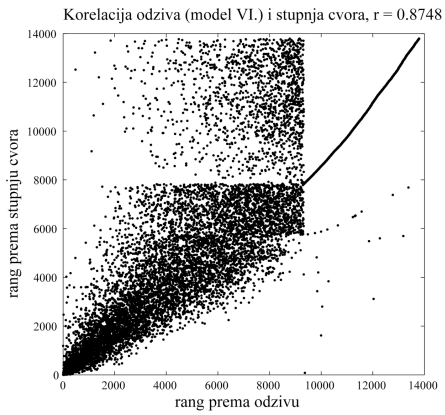


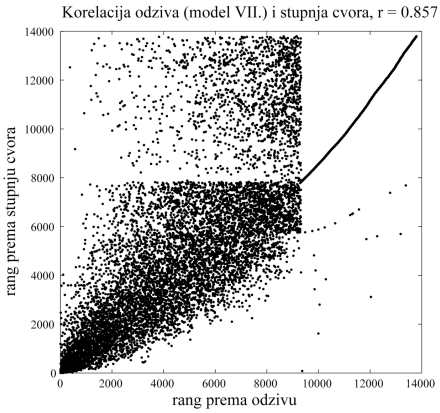


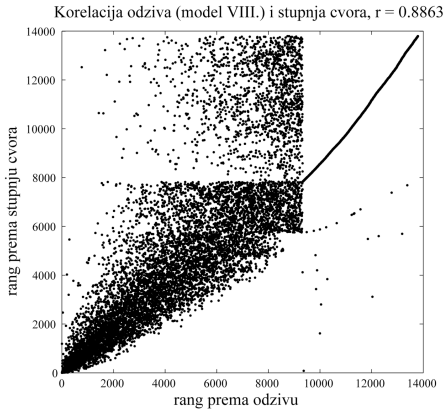


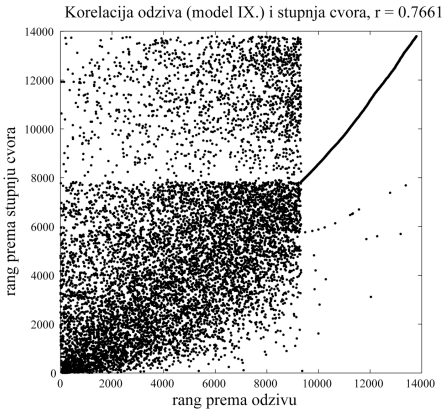




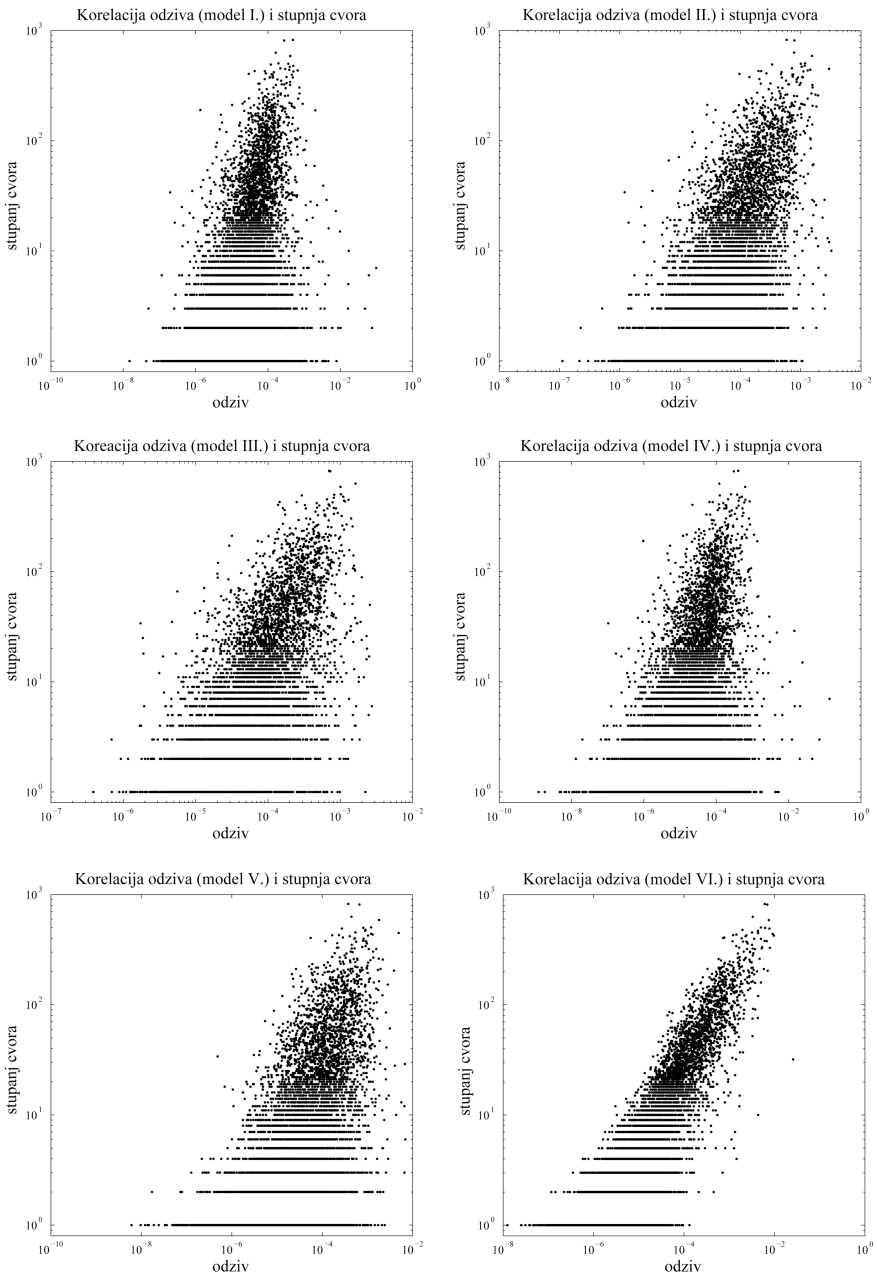


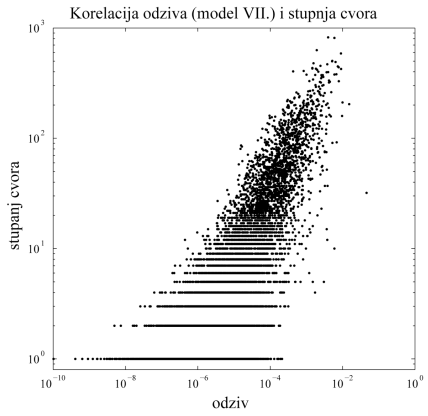

Korelacija odziva (model VII.) i stupnja cvora, r = 0.8571
rang prema stupnju cvora
rang prema odzivu
0
2000
4000
6000
8000
10000
12000
14000

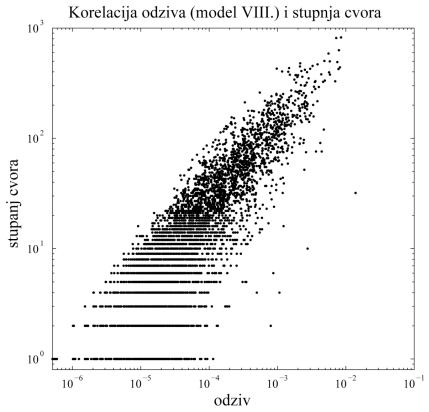

Korelacija odziva (model VIII.) i stupnja cvora, r = 0.8863
rang prema stupnju cvora
rang prema odzivu
0
2000
4000
6000
8000
10000
12000
14000

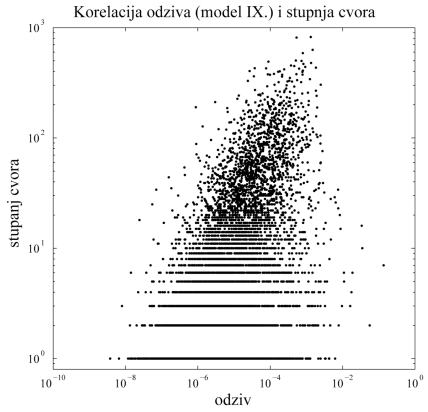

Korelacija odziva (model IX.) i stupnja cvora, r = 0.7661
rang prema stupnju cvora
rang prema odzivu
0
2000
4000
6000
8000
10000
12000
14000

Usporedba vrijednosti dobivenih odzivom i stupnjem čvora

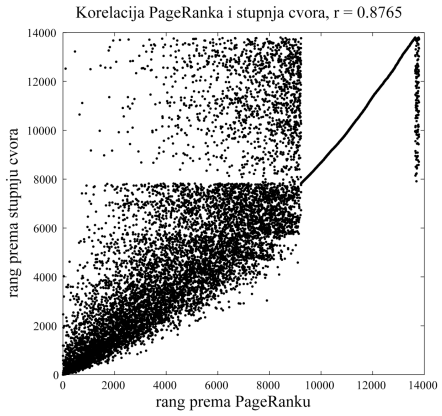


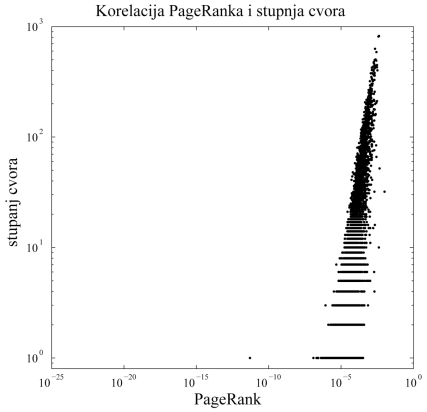






Usporedba PageRanka i stupnja čvora





6 Diskusija

Kao glavni indikator kakvoće korelacije uzet će se koeficijent korelacije $\boldsymbol{r}$. On je izračunat za sve usporedbe u kojima je korelirana uspješnost rangiranja po pojedinoj mjeri. Apsolutni iznos koeficijenata korelacije, ipak treba uzeti s rezervom, jer veliki doprinos daju čvorovi koji su izolirani (ne posjeduju ni jednu vezu). Takvih čvorova ima par tisuća i rangirani su na zadnje pozicije prema svim mjerama. Kako je njihov broj stalan i jednako pridonose koeficijentu korelacije u svim usporedbama njihov utjecaj na poredak mjera vjerojatno nije značajan.

Devet predloženih modela odziva poredani su po uspješnosti korelacije s PageRankom:

MODEL ODZIVA	KORELACIJA S PAGERANKOM
1. model VIII.	0.9584
2. model VI.	0.9533
3. model VII.	0.9412
4. model V.	0.8577
5. model II.	0.8462
6. model III.	0.8344
7. model IV.	0.8204
8. model IX.	0.8165
9. model I.	0.7821

Primjećuje se da postoje tri modela odziva koji daju izvrsne rezultate s koeficijentima koralacije većim od 0.94. Ostali modeli imaju korelaciju manju od 0.86 što se u ovom kontekstu ocijenjuje nezadovoljavajućim. Razlog za to je korelacija PageRanka i stupnja čvora koja iznosi 0.8765. Od kvalitetne mjere odziva očekujemo barem da bude bolja od toga. To ne znači naravno da je korelacija s PageRankom jedino što nas zanima, ali u ovom slučaju PageRank smatramo dovoljno relevantnim da posluži kao pouzdana mjera za usporedbu.

Navodimo još jednom tri najuspješnija modela:

VI. model

$$\boldsymbol{R(u) = \frac{\sum_{i=1}^{N} C_i^2(u)}{\sum_{i=1}^{N} P_i(u)} + \sum_{i=1}^{N} C_i(u)}$$

VII. model

$$\boldsymbol{R(u) = \sum_{i=1}^{N} C_i^2(u)}$$

VIII. model

$$\boldsymbol{R(u) = \frac{|C(u)|^2}{|P(u)|} + |C(u)|}$$

Kako se pokazalo, veliku ulogu u kvaliteti korelacije s PageRankom pokazuju modeli koji naglasak stavljaju na komentare. Takvi su upravo najbolja četiri rangirana modela. Iznimka je model IV. koji je rangiran relativno nisko.

Ako se provjeri korelacija sa stupnjem čvorova dobiva se poredak:

MODEL ODZIVA	KORELACIJA SA STUPNJEVIMA
1. model VIII.	0.8863
2. model VI.	0.8748
3. model VII.	0.8571
4. model III.	0.7873
5. model II.	0.7850
6. model V.	0.7846
7. model IX.	0.7661
8. model IV.	0.7550
9. model I.	0.7259

Dobiveni rangovi gotovo su identični onima za PageRank - prva tri modela (VIII., VI. i VII.) zauzimaju isti poredak. Maksimalna korelacija nije puno veća od 0.87 koliko iznosi korelacija stupnjeva s PageRankom. Očito sličnosti u korelaciji mjere odziva i PageRanka u usporedbi sa stupnjem čvora nisu u potpunosti slučajne. To je na neki način i razumljivo uzme li se u obzir da obje mjere pokušavaju kvantificirati "važnost" čvora - PageRank preko globalne strukture veza, a odziv prema karakteristikama samog čvora.

Ovdje prezentirani rezultati su obećavajući, no potrebna su daljnja istraživanja kako bi se preciznije utvrdila opravdanost korištenja mjere odziva i njena potencijalna primjena.

7 Daljnji rad

Identifikacija ključnih igrača zanimljiva je tema u okviru koje postoji još puno dodatnih problema za istraživanje. U ovom radu dotaknut je samo dio njih. Neki od prijedloga za daljnji rad su:

- Ispitati korelaciju mjere odziva s ostalim poznatijim mjerama centralnosti - Freemanovom međupoloženošću prije svega.
- Provjeriti kakav utjecaj na odziv i ostale mjere centralnosti imaju neregistrirani korisnici.
- Kako izolirani čvorovi uvelike doprinose apsolutnom iznosu korelacije trebalo bi provesti analizu u kojoj bi se njih izbacilo iz proračuna.
- Ograničiti vremenski period iz kojeg se uzimaju podatci za mreža blogova. Trenutno u analizu ulaze postovi i komentari iz vremenskog raspona od par godina (tj. od početka djelovanja `bloger.hr` servisa) pa mnoge veze sasvim sigurno više nisu aktualne.
- Zasad jedina karakteristika postova i komentara koja se uzima u obzir je njihov broj i duljina u znakovima. Pri tome su blogovi koji sadrže manje teksta a više multimedije - fotografije na primjer, u podređenoj poziciji. Trebalo bi pokušati dohvatiti informaciju o broju i vrsti multimedije na blogovima, te definirati mjeru odziva koja bi i takve podatke uzela u obzir.
- Generalizirati definiciju mjere odziva kako bi ona bila primjenjiva i na mreže iz drugih domena. Prije svega, na ostale vrste socijalnih mreža u kojima se može izmjeriti lokalni intezitet međusobne povezanosti. Veliki potencijal za dobivanje kvalitetnih modela mreža imaju web servisi koji se bave povezivanjem i suradnjom većeg broja korisnika[10] kao što su Facebook, Last.fm, MySpace...

[10] Općeniti naziv za takvu filozofiju dizajna je Web 2.0.

Zaključak

O pojmovima kao što su važnost ili autoritet razmišljamo prije svega u subjektivnim okvirima. S druge strane, struktura mreže nešto je što je lako formalno definirati iako za centralnost u kontekstu mreža još uvijek ne postoji jedinstvena definicija. Jedina pretpostavka koja vrijedi za sve mjere je da je centralnost svojstvo na razini čvorova [1] u smislu da su upravo čvorovi ti koji su više ili manje centralni.

No kako definirati jedinstveno svojstvo centralnosti isključivo na temelju strukture mreže koje će biti primjenjivo na široki spektar problema, ostaje otvoreno pitanje. Opravdanost takvog pristupa, barem s praktičnog stajališta, tek treba utvrditi. Vjerojatno nije daleko od istine tvrdnja da je svojstvo centralnosti nešto što je inherentno kontekstu domene kojoj mreža pripada. Bez odgovarajuće interpretacije, općenitost strukturalnog pristupa nije od prevelike koristi. Primjerice, definiranje međupoloženosti toka ima smisla samo ako je u potpunosti jasno što je to tok i kako se on manifestira u mreži. U suprotnom, takva mjera rangira čvorove po kriteriju, koji nam u kontekstu u kojem se nalazimo, ne znači ništa.

Rezultati ovog istraživanja pokazuju da je moguće definirati mjeru koja nije temeljena na strukturalnom pristupu, a koja relativno dobro korelira sa ostalim mjerama. Pogotovo je obećavajuća korelacija nekih modela s PageRankom - algoritmom temeljenom na globalnom strukturalnom pristupu koji danas ima uspješnu komercijalnu primjenu. Razne varijante odziva mogle bi biti primjenjive i na mreže iz drugih domena, prije svega na mreže socijalnih interakcija.

Važnost ili autoritet nije nešto što proizlazi iz strukture mreže već upravo suprotno - to je aktivno svojstvo agenata koji sudjeluju u stvaranju mrežne strukture. Ako se iz strukture *a posteriori* i može zaključiti nešto o važnosti i autoritetu pojedinog čvora, to je i dalje samo posljedica međudjelovanja agenata koji su doveli do takve strukture. Pristup u kojem se izbjegava modeliranje cijele strukture mreže i koji dovodi do kvalitetnih zaključaka samo uz pomoć lokalnih karakteristika čvora, mogao bi dovesti do razvitka novih metoda u istraživanju centralnosti. Takve metode mogle bi se na koncu pokazati učinkovitije ili barem jednako dobre kao one bazirane na strukturalnom pristupu.

Literatura

[1] Stephen P. Borgatti, Martin G. Everret: *A Graph-theoretic perspective on centrality*, Social Networks 28 (2006) 466-484

[2] Jon M. Kleinberg: *Authoritative Sources in a Hyperlinked Environment*, Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, 1998.

[3] Sergey Brin, Larry Page: *The PageRank Citation Ranking: Bringing Order to the Web*, 29. siječnja 1998.

[4] D. S. Sade: *Sociometrics of macaca mulatta III: N-path centrality in grooming networks*, Social Networks 11, 273-292

[5] L. C. Freeman: *The gatekeeper, pair-dependency and structural centrality*, Quality and Quantity 14, 585-592

[6] N. E. Friedkin: *Theoretical foundations for centrality measures*, American Journal of Sociology 96, 1478-1504

[7] A. L. Barabási, R. Albert: *Emergence of scaling in random networks*, Science 286, 509.

[8] L. Wall, T. Christiansen, J. Orwant: *Programming Perl*, O'Reilly, 2000.

Sažetak

U okviru ovog rada predstavljena je osnovna klasifikacija mjera centralnosti temeljenih na strukturalnom pristupu. Detaljnije su opisane Sadeova k-centralnost (eng. *Sade's k-path centrality*), Freemanova međupoloženost (eng. *Freeman's betweenness*) i PageRank. Nadalje, predložena je nova mjera centralnosti koja se temelji na svojstvu *odziva* - omjeru broja ili duljine komentara i postova. Razvijeno je devet modela mjere odziva.

Testiranje i analiza provedena je na modelu mreže blogova u kojem blogeri i njihovi blogovi imaju ulogu čvorova, a njihovi međusobni komentari ulogu veza. Podatci su dobiveni sa servisa `www.bloger.hr` koji djeluje od 2004. i do danas broji oko 13000 registriranih korisnika.

Za implementaciju su odabrane dvije strukturalne mjere - stupanj čvora (poseban slučaj Sadeove k-centralnosti u kojoj je $\boldsymbol{k = 1}$) i PageRank. Prva se temelji na lokalnoj strukturi mreže - broju veza koje čvor posjeduje, a druga na globalnoj koja ovisi o strukturi cijele mreže i rasporedu svih veza unutar nje. Također je implementirano i svih devet modela odziva, te su dobiveni rezultati uspoređeni. Za svaku usporedbu izračunat je koeficijent korelacije.

Pokazalo se da neki modeli odziva izvrsno koreliraju s odabranim strukturalnim mjerama. Pogotovo je obećavajuća korelacija nekih modela s PageRankom - algoritmom temeljenom na globalnom strukturalnom pristupu koji danas ima uspješnu komercijalnu primjenu. Dobiveni rezultati sugeriraju da je moguće i opravdano, definirati kvalitetne mjere koje svoje uporište neće imati u globalnoj strukturi mreže, već u lokalnim karakteristikama samih čvorova i njihovih veza.

Naslov rada: Identifikacija ključnih igrača u kompleksnim mrežama blogova

Ključne riječi: ključni igrači, centralnost, odziv, međupoloženost, PageRank

Abstract

The basic classification of centrality measures based on structural approach has been presented within the scope of this thesis. Sade's k-path centrality, Freeman's betweenness and PageRank are presented in detail. Also, a new measure of centrality based on the property of *response* is proposed. Nine models are developed that differ in the ratio between the length and count of comments and posts.

Testing and analysis is performed on the network model based on blogs in which blogers have the role of nodes and their comments role of connections. Data has been acquired from www.bloger.hr which is active from 2004. and counts around 13000 registered blogers to this day.

Two structural measures are chosen for implementation - degree of the node (special case of Sade's k-path centrality where k = 1) and PageRank. The first one is based on the local network structure - number of undirected connections that node has, and the other one on the global network structure that depends on the arrangement of all connections in it. The nine models of response are also implemented and results are compared with other measures. Correlation coefficient is calculated for every comparison.

It's shown that some of the response models show excellent correlation with other structural measures. Especially promising is the correlation between some models and the PageRank - the algorithm that's based on global structural approach and that has successful commercial application. The results suggest that definition of measures that are not based solely on the local or global structure of network, but on the local properties of the nodes themselves instead, is possible and justified.

Thesis title: Identification of key players in complex networks of blogs

Keywords: key players, centrality, response, betweenness, PageRank

Dodatak 1

Uz ovaj završni rad priložen je CD s potpunim tekstom rada i programskim kodovima za dohvaćanje podataka i analizu mreže blogova. Na CD-u se nalaze i tekstualne datoteke u koje su spremljeni rezultati analize. Bazi podataka koju je koristio autor, nije moguće pristupiti, no kodovi su u potpunosti funkcionalni. Sadržaj CD-a je sljedeći:

```
..\Tekst\

    ZR[2008]Piskorec_Matija.pdf

..\Kodovi\

    blogspace_final.pl
    BlogspaceLib.pm
    parsing_information.txt
    ..\Kodovi\database\

        create_blogspace.sql

    ..\Kodovi\network\

        calculate_degree.pl
        calculate_pagerank.pl
        calculate_response.pl
        compare_mesures.pl
        create_network.pl
        network_parameters.pl
        ..\Kodovi\network\statistics\

            cumulative_distribution.txt
            degree_vs_pagerank.txt
            degree_vs_pagerank_values.txt
            distribution_backlinks.txt
            distribution_degree.txt
            distribution_forwardlinks.txt
            distribution_response.txt
            response_vs_degree.txt
            response_vs_degree_values.txt
            response_vs_pagerank.txt
            response_vs_pagerank_values.txt
```