Rudarenje podataka o znanstvenoj suradnji iz baze google scholar

Murljačić, Elena

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Elena Murljačić

RUDARENJE PODATAKA O ZNANSTVENOJ SURADNJI IZ BAZE GOOGLE SCHOLAR

Diplomski rad

Voditelj rada:
prof. dr. sc. Luka Grubišić

Zagreb, rujan, 2020.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Zahvaljujem prof.dr.sc. Luki Grubišiću za sve savjete i pomoć tijekom pisanja ovog rada. Zahvaljujem cijeloj svojoj obitelji, dečku i prijateljicama za strpljenje i podršku tijekom studija. Zahvaljujem svojim kolegicama koje su sa mnom dijelile sve tuge i radosti ovog perioda života. Na kraju zahvaljujem Onome koji mi je darovao talente i snagu da ovo mogu uspješno savladati.

Sadržaj

Uvod

Ako razmišljamo o definiciji što bi nešto činilo jednostavnim ili kompleksnim, ne možemo precizno izraziti. No, obično imamo intuiciju za prepoznati kada je nešto jednostavno ili kompleksno. Svuda oko nas nalaze se kompleksni sustavi. Naprimjer, WWW (World Wide Web) je skup povezanih web stranica, društvene mreže su sustavi ljudi, a ljudsko tijelo je kompleksni sustav mišića, vlakna, proteina, stanica itd. Cijeli naš svemir je jedan kompleksni sustav. Mreže su skupovi čvorova (koje nekada zovemo i vrhovi) koji su povezani bridovima te se u matematici zovu grafovi. Kompleksne mreže su grafovi kojima želimo modelirati neke sustave iz različitih domena našeg svijeta, tako da čvorovi predstavljaju elemente sustava, a bridovi njihovu relaciju. Neki od primjera bili bi sljedeći: socijalne mreže (npr. osobe su povezane ako su prijatelji,), informacijske (npr. WWW - web stranice koje su povezane poveznicama), biološke (npr. vrste su povezane ako ima predator-plijen vezu).
Rudarenje podataka (engl. data mining) je proces otkrivanja anomalija, uzoraka i korelacija u velikim bazama podataka pritom upotrebljavajući alate statistike, strojnog učenja ili slično. Koristeći rudarenje podataka, pretvaramo podatke u korisne informacije.

U ovom radu ćemo analizirati jednu socijalnu kompleksnu mrežu - mrežu suradnje među znanstvenicima. Podaci za izradu mreže preuzeti su sa servisa Google Scholar, koji je široko rasprostranjena baza podataka o znanstvenim radovima. Rad se sastoji od pet poglavlja. U prva 3 poglavlja obradit ćemo osnove teorije kompleksnih mreža. U četvrtom poglavlju analizirat ćemo kompleksnu mrežu, otkrivati uzorke ili anomalije i aproksimirati mrežu s nekoliko matematičkih modela. U petom poglavlju sažet ćemo dobivene rezultate i donijeti zaključak.

Poglavlje 1

Teorija grafova

Da bismo matematički proučili kompleksne mreže, modeliramo ih grafovima - podatke možemo reprezentirati vrhovima grafa, a veze među podacima možemo reprezentirati bridovima. U ovom poglavlju iznosimo pojmove iz teorije grafova koji će nam koristiti za razumijevanje svojstva i modela kompleksnih mreža. Definicije iz ovog poglavlja preuzete su iz [9] i [10].

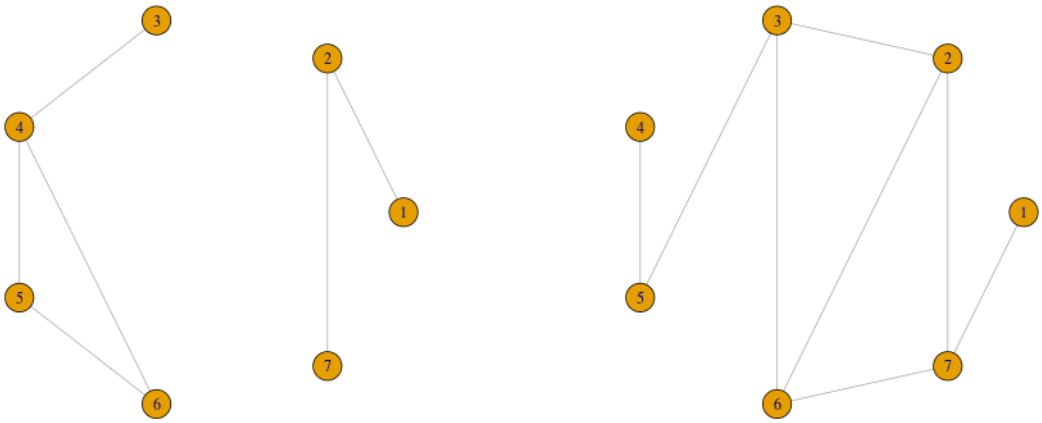
1.1 Osnovni pojmovi iz teorije grafova

Definicija 1.1.1. *(Neusmjereni)* ***Graf*** *je uređeni par $G = (V, E)$, gdje je $V = V(G)$ neprazan skup čije elemente nazivamo* ***vrhovima****, a $E = E(G)$ je familija dvočlanih podskupova od V koje nazivamo* ***bridovima****.*
Pritom vrhove u i v koji su pridruženi bridu e nazivamo krajevima brida e, a brid čiji se krajevi podudaraju nazivamo petljom.

Definicija 1.1.2. ***Usmjereni graf*** *je uređeni par $G = (V, E)$, gdje je $V = V(G)$ neprazan skup čije elemente nazivamo* ***vrhovima****, a $E = E(G)$ je skup uređenih parova elemenata iz V, čije elemente zovemo* ***usmjerenim bridovima****.*

Definicija 1.1.3. ***Težinski graf*** *G je graf kojem je pridružena funkcija $f : E \to \mathbb{R}$, koja svakom bridu iz grafa pridružuje njegovu težinu. Tu funkciju nazivamo* ***težinskom funkcijom****.*

Definicija 1.1.4. *Kažemo da su vrhovi $u, v \in V$ grafa $G = (V, E)$* ***susjedni*** *ako postoji brid e kojem su oni krajevi.U slučaju usmjerenog grafa pišemo $e = (u, v)$, a u slučaju neusmjerenog $e = \{u, v\}$. Pritom kažemo da je brid e* ***incidentam*** *s vrhovima u i v. Općenito pišemo $e = uv$.*



Slika 1.1: Lijevo: graf s dvije komponente povezanosti; Desno: povezani graf

Definicija 1.1.5. ***Stupanj vrha*** *v u neusmjerenom grafu G je broj bridova grafa G koji su incidentni s v. Svaka petlja računa se kao dva brida.*
Izolirani vrh *je vrh stupnja 0.* ***Ulazni stupanj vrha*** *v u usmjerenom grafu G je broj bridova grafa G oblika $e = (x, v)$.*
Analogno se definira ***izlazni stupanj vrha****.*

Definicija 1.1.6. ***Šetnja*** *u grafu $G = (V, E)$ je niz $(v_0, e_1, v_1, e_2, v_2, ..., e_n, v_n)$, gdje je $e_i \in E$ brid $v_{i-1}v_i$ za $i = 1, ..., k$.*
Put *je šetnja u kojoj su svi vrhovi različiti osim eventualno prvog i zadnjeg.*
Duljina puta *je broj bridova u nizu za bestežinske grafove, odnosno suma vrijednosti težinske funkcije svih bridova u nizu za težinske grafove.*
Kažemo da su dva vrha ***povezana*** *ako postoji put između njih.*
Kažemo da je graf ***povezan*** *ako su svi njegovi vrhovi povezani.*

Definicija 1.1.7. *Kažemo da je graf $G' = (V', E')$* ***podgraf*** *grafa $G = (V, E)$ ako je $V' \subseteq V$ i $E' \subseteq E$.*

Definicija 1.1.8. ***Komponenta povezanosti*** *grafa je maksimalni povezani podgraf grafa.*

Definicija 1.1.9. *Neka je $G = (V, E)$ graf i $n = |V|$ broj vrhova. Neka je $V = v_1, v_2, ..., v_n$.* ***Matrica susjedstva*** *grafa G je matrica $\mathbf{A} \in M^{n \times n}$ čiji su elementi dani s:*

$$A_{ij} = \begin{cases} 1 & \textit{postoji brid koji povezuje vrhove i,j} \\ 0 & \textit{inače}. \end{cases} \tag{1.1}$$

Primjer 1.1.10. *Sljedeća matrica je matrica susjedstva za lijevi graf sa slike 1.1*

$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Poglavlje 2

Svojstva kompleksnih mreža

U grafovima koji prestavljaju kompleksne mreže možemo mjeriti određene veličine. Te veličine koristimo kako bismo proučili svojstva kompleksnih mreža. U ovom poglavlju iznijet ćemo neke veličine iz područja teorije grafova koje koristimo pri određivanju svojstava različitih kompleksnih mreža.

2.1 Svojstvo malog svijeta

U eksperimentu Stanleya Milgrama iz 1960-ih, sudionicima je bilo dano pismo s imenom neke druge osobe te su sudionici zamoljeni da pismo predaju nekom poznaniku. Većina pisama je bilo izgubljeno, no otprilike četvrtina pisama je stiglo na odredište i pisma su u prosjeku prošla samo šestoro ljudi. Taj rezultat je prvi pokazatelj postojanja svojstva malog svijeta (engl. small-world effect), činjenice da je u većini mreža, većina parova vrhova povezana putem kratke duljine [7].

Definicija 2.1.1. *Neka je $G = (V, E)$ graf. Označimo s d_{ij} duljinu najkraćeg puta od vrha i do vrha j. Ako je $n = |V|$ broj vrhova grafa G tada*

$$\ell = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij} \tag{2.1}$$

označava prosječnu duljinu najkraćeg puta između para vrhova u grafu.

Definicija 2.1.2. *Neka je ℓ prosječna duljina najkraćeg puta i $n = |V|$ broj vrhova u mreži. Za mrežu kažemo da ima **svojstvo malog svijeta** ako vrijedi:*

$$\ell \sim O(log(n)) \tag{2.2}$$

Definicija 2.1.3. *Neka je $G = (V, E)$ graf. Označimo s d_{ij} duljinu najkraćeg puta od vrha i do vrha j.* ***Promjer grafa*** *G definiramo kao:*

$$d_G = \max_{i,j} d_{ij} \tag{2.3}$$

2.2 Tranzitivnost

U mnogim mrežama nalazimo sljedeće svojstvo: ako je vrh i povezan s vrhom j i vrh j s vrhom k, tada postoji velika vjerojatnost da je vrh i povezan s vrhom k. To svojstvo nazivamo **tranzitivnost** (ili svojstvo klasteriranja). U socijalnim mrežama (kakve su i mreže suradnje) to je ekvivalentno sljedećem: ako su osobe i i j prijatelji, te osobe j i k prijatelji, tada će vjerojatno osobe i i k također biti prijatelji.
Da bismo promotrili tranzitivnost mreže, definirat ćemo koeficijent klasteriranja za pojedini vrh te prosječni koeficijent klasteriranja za graf.

Definicija 2.2.1. ***Koeficijent klasteriranja*** *C_i za vrh i definiramo kao*

$$C_i = \frac{n^i_\triangle}{n^i_{triples}} \tag{2.4}$$

gdje $n^i_\triangle$ označava broj "trokuta" (podgrafa s 3 vrha i 3 brida) povezanih s vrhom i, te $n^i_{triples}$ označava broj povezanih trojki vrhova u kojima je vrh i središnji vrh. Za vrhove stupnja 0 ili 1, definiramo $C_i = 0$.
Prosječni koeficijent klasteriranja *C definiramo kao*

$$C = \frac{1}{n} \sum_i C_i \tag{2.5}$$

2.3 Centralnost

U kompleksnim mrežama možemo razmatrati koliko je neki vrh bitan u mreži, npr. koliko je neka osoba bitna u socijalnoj mreži. Mjeru bitnosti nekog vrha za određenu kompleksnu mrežu nazivamo **centralnost**.
Centralnost možemo promatrati na različite načine. Najjednostavnije, možemo promatrati stupanj pojedinog vrha da bismo odredili koliko je centralan.

Definicija 2.3.1. ***Centralnost stupnja*** *za vrh i definiramo kao*

$$c_d(i) = \frac{k_i}{n-1} \tag{2.6}$$

gdje k_i označava stupanj vrha i, te je n broj vrhova grafa.

Centralnost nekog elementa mreže možemo promatrati i preko centralnosti njegovih susjeda. Ako je element mreže povezan s izrazito centralnim elementima, on sam će biti više centralan [4]. To se matematički može zapisati kao sustav jednadžbi

$$x_i = \frac{1}{\lambda} \sum_{j=1}^{n} A_{ij} x_j \tag{2.7}$$

gdje je $\mathbf{A} = (A_{ij})$ matrica susjedstva, te λ neka konstanta.

Sustav se može zapisati na sljedeći način:

$$\mathbf{A} = \lambda \mathbf{x} \tag{2.8}$$

iz čega vidimo da se radi o svojstvenom problemu. Uz uvjet da su vrijednosti iz svojsvetnog vektora $\mathbf{x}$ pozitivne, prema [4] slijedi da nam samo svojstveni vektor uz najveću svojstvenu vrijendost daje odgovarajuću mjeru centralnosti.

Definicija 2.3.2. *Neka je $G = (V, E)$ graf i $\mathbf{A} = (A_{ij})$ matrica susjedstva.* ***Centralnost svojstvenog vektora*** *vrha i je i-ta komponenta svojstvenog vektora pridruženom najvećoj svojstvenoj vrijendosti λ matrice $\mathbf{A}$.*

Kod nekih vrhova može se dogoditi da imaju veliki stupanj, što znači da su centralniji u smislu centralnosti stupnja, ali imaju veću prosječnu udaljenost od ostalih vrhova. Zato promatramo **centralnost blizine** vezanu uz udaljenost među vrhovima, gdje nam manje vrijednosti označavaju malenu udaljenost od ostalih vrhova.

Definicija 2.3.3. *Neka je $G = (V, E)$ graf te $i \in V$ neki vrh. Za vrh $j \in V$ označimo s d_{ij} duljinu najkraćeg puta od vrha i do vrha j. Tada s*

$$c_c(i) = \frac{n-1}{\sum_{j \neq i} d_{ij}} \tag{2.9}$$

definiramo ***centralnost blizine*** *za vrh i.*

2.4 Gigantska komponenta

Komponente povezanosti grafa posebno su značajne u promatranju širenja epidemije u mreži [4]. Naprimjer, ako promatramo neko selo ili grad kao mrežu gdje vrhovi predstavljaju ljude, a bridovi predstavljaju poznanstvo, tada počinjući od neke osobe možemo promatrati širenje epidemije. Ako je osoba dio neke manje komponente povezanosti, epidemija se neće puno proširiti. No, ako je osoba dio velike komponente povezanosti koja obuhvaća puno veći dio mreže, epidemija će se vjerojatno proširiti na čitav grad.

Definicija 2.4.1. ***Gigantska komponenta*** *(engl. giant component) grafa je skup vrhova koji pripadaju nekoj komponenti povezanosti čija je veličina $O(n)$, gdje je n broj vrhova u grafu.*

Većina mreža, iz stvarnog svijeta ili modela, ima najviše jednu gigantsku komponentu u grafu, dok su ostale komponente povezanosti veličine $O(1)$ [4].

2.5 Distribucija stupnja

Definicija 2.5.1. ***Distribucija stupnja*** *je funkcija distribucije $P(k)$ koja označava vjerojatnost da je nasumnično odabran vrh supnja k. Alternativno, $P(k)$ je udio vrhova stupnja k u mreži.*
Srednji stupanj (prvi moment distribucije) je

$$\langle k \rangle = \sum_{k=0}^{\infty} kP(k) \tag{2.10}$$

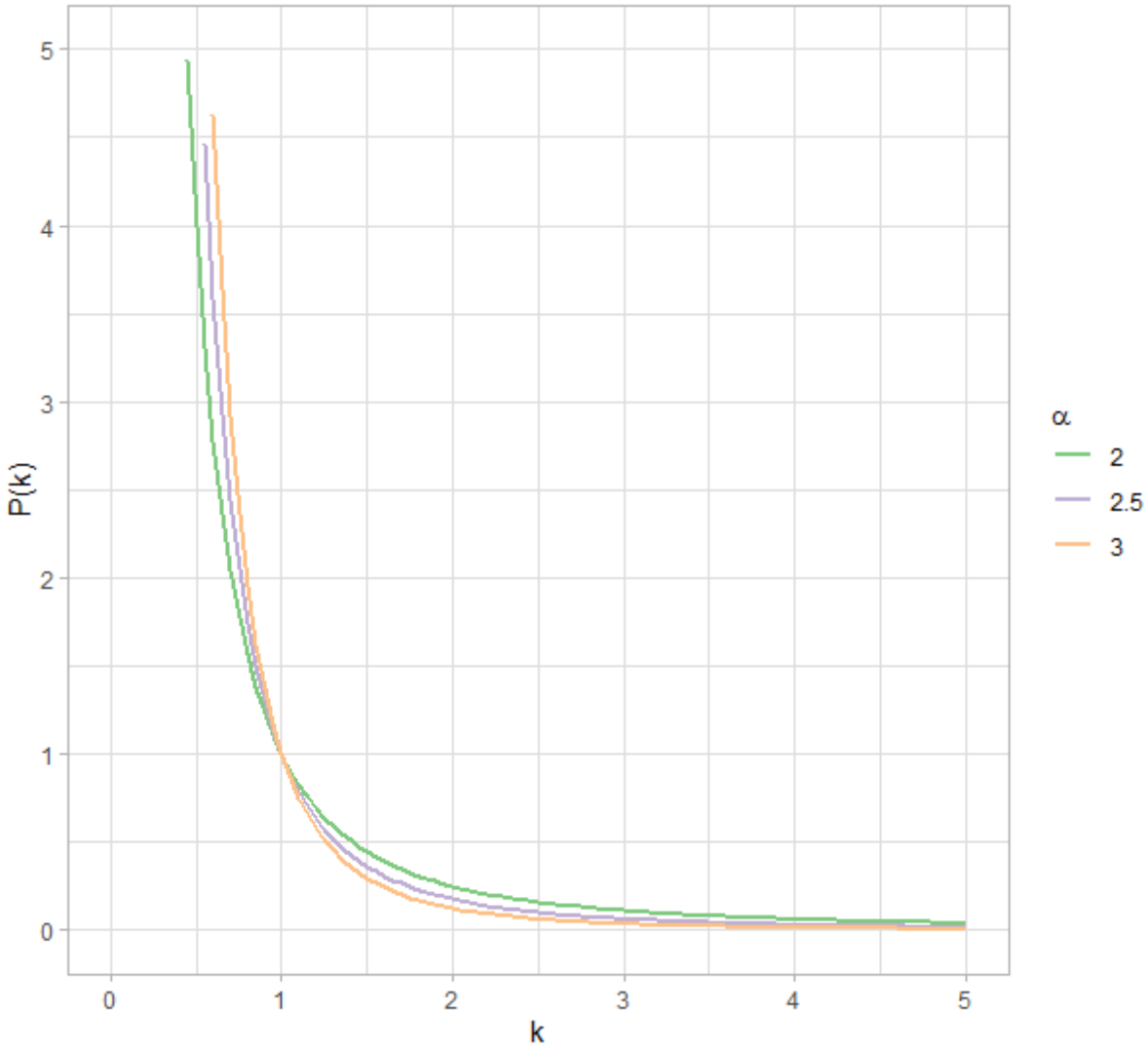
n-ti moment distribucije je

$$\langle k^n \rangle = \sum_{k=0}^{\infty} k^n P(k) \tag{2.11}$$

Definicija 2.5.2. ***Kumulativna distribucija stupnja*** *je funkcija distribucije $F(k)$ koja označava vjerojatnost da je nasumnično odabran vrh stupnja većeg ili jednakog k. Pišemo:*

$$F(k) = \sum_{k'=k}^{\infty} P(k') \tag{2.12}$$

Mreže iz "stvarnog svijeta" najčešće su karakterizirane iskrivljenim distribucijama stupnja s "teškim repovima", što znači da u mreži ima mnogo vrhova većeg stupnja kojh ima jednako mnogo. Najčešća takva distribucija je distribucija zakona potencija (engl. power-law distribution), $P(k) \sim k^{-\alpha}, \alpha = const.$ .Najčešća vrijednost parametra α je između 2 i 3 [4]. U tom slučaju kumulativna distribucija bila bi: $F(k) \sim \sum_{k'=k}^{\infty} k'^{-\alpha} \sim k^{-(\alpha-1)}$ [7]. Često je korisnije crtati kumulativnu distribuciju jer pri crtanju histograma za distribuciju možemo izgubiti neke razlike među podacima koji spadaju u isti stupac.



Slika 2.1: Distribucija zakona potencija

Poglavlje 3

Osnovni modeli kompleksnih mreža

U ovom poglavlju opisat ćemo matematičke definicije za neke od osnovnih modela kompleksnih mreža kojima se želi pokušati opisati i analizirati ponašanje mreža iz stvarnog svijeta. Razmatrat ćemo neusmjerene grafove bez težina.

3.1 Model slučajnog grafa

Najpoznatiji model kompleksne mreže svakako je model slučajnog grafa, kojeg još nazivamo Poissonov slučajni graf. Taj model prvi su neovisno prikazali Rapoport te Erdős i Rényi [7]. Ovaj model je prilično jednostavan, ali neadekvatan za opisivanje nekih svojstava mreža iz stvarnog svijeta.

Poissonov slučajni graf

Neka je n broj vrhova i $p \in [0, 1]$ vjerojatnost povezivanja dvaju vrhova bridom. Graf $G_{n,p}$ konstruiramo tako da svaki od $\frac{n(n-1)}{2}$ bridova generiramo s vjerojatnošću p, odnosno ne generiramo s vjerojatnošću $1 - p$. Svojstva ovakvog modela određena su za dovoljno veliki n [7]. Kako se svaki brid generira neovisno, vjerojatnost da će vrh biti stupnja k dana je s

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{3.1}$$

Srednji stupanj grafa je

$$\langle k \rangle = (n-1)p \sim np. \tag{3.2}$$

Za dovoljno veliki n i fiksni k, distribucija teži u Poissonovu distribuciju,

$$P(k) = \frac{\langle k \rangle e^{-\langle k \rangle}}{k!}. \tag{3.3}$$

Vjerojatnost da su dva vrha povezana je p bez obzira na to imaju li zajedničkog susjeda. Stoga koeficijent klasteriranja iznosi p te teži u 0 kada $n \to \infty$.
Prema [9], može se pokazati da vrijedi:

$$\ell \simeq \frac{\log n}{\log \langle k \rangle} \tag{3.4}$$

Iz toga slijedi da model ima svojstvo malog svijeta.

Generalizirani slučajni graf

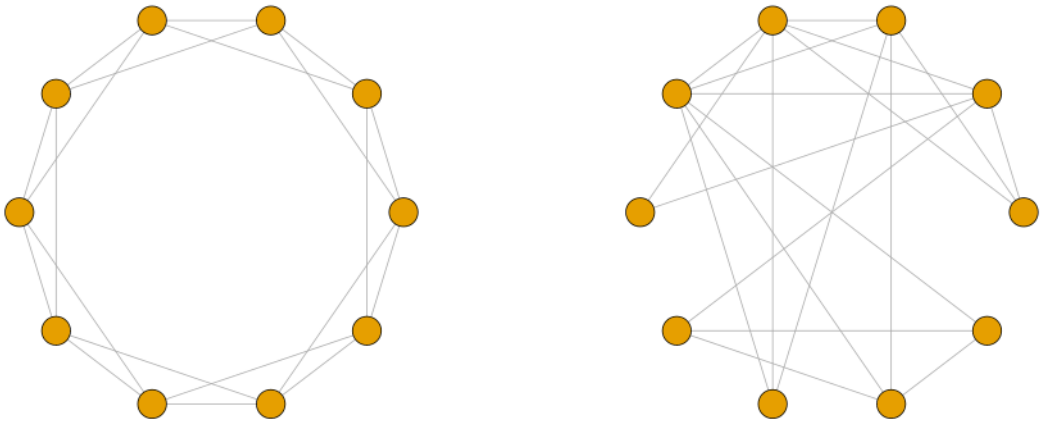
Da bismo u model slučajnog grafa ukomponirali realističnije distribucije stupnja te tako dobili adekvatniji model, razmatramo **konfiguracijski model**. Specificirat ćemo distribuciju stupnja $P(k)$. Iz te distribucije odabiremo skup od n vrijednosti stupnjeva k_i, za $i = 1, ..., n$, gdje je n broj vrhova u grafu. Zatim u skladu s prethodnim na slučajan način dodamo bridove.
Može se pokazati da će koeficijent klasteriranja uobičajeno težiti nuli za velike grafove, ali za neke distribucije "teškog repa" koeficijent klasteriranja neće biti zanemariv [7].
Može se pokazati da prosječni najkraći put ima logaritamsku ovisnost o veličini grafa, stoga ovaj model ima svojstvo malog svijeta [9].

3.2 Model malog svijeta

Za razliku od modela slučajnog grafa, model malog svijeta daje nam graf visoke tranzitivnosti, to jest većeg koeficijenta klasteriranja.
Neka je n broj vrhova, $p \in [0, 1]$ i $\langle k \rangle = 2m$ željeni prosječni stupanj grafa. Prema [9], graf konstruiramo na sljedeći način:

- svaki od n vrhova povežemo simetrično s 2m najbližih susjeda
- sa svakim od m desnih susjeda ostajemo povezani s vjerojatnošću $1 - p$, odnosno nasumnično odaberemo neki drugi vrh s vjerojatnošću p

Na slici 3.2 prikazani su primjeri grafova konstruirani na takav način.
Svojstva modela malog svijeta ovise o p. Za $p = 0$, koeficijent klasteriranja biti će velik [7]. Za isti p, neće vrijediti svojstvo malog svijeta zbog velike dužine prosječnog najkraćeg puta. Za male vrijednosti p, pojavit će se "prečaci" između vrhova te će se ta vrijednost smanjiti. Za $p = 1$, zbog nasumničnosti će koeficijent klasteriranja biti vrlo malen. Graf će se također zbog nasumničnosti približiti slučajnom grafu što znači da će prosječni najkraći put imati logaritamsku ovisnost o veličini grafa i model će imati svojstvo malog svijeta [7].
Može se izračunati distribucija stupnja te pokazati da je "lakog repa" [9].



Slika 3.1: Lijevo: $n = 10$, $m = 2$, $p = 0$; Desno: $n = 10$, $m = 2$, $p = 0.3$

3.3 Model eksponencijalnih slučajnih grafova

Generalizirani slučajni grafovi uspješno su riješili jedan od problema Poissonovih slučajnih grafova, nerealnu distribuciju stupnja. Pomoću ovog modela možemo uspješno modelirati neka druga svojstva, kao što je tranzitivnost.
Neka je ϵ_i skup mjerljivih svojstva grafa (npr. broj bridova, broj vrhova određenog stupnja, broj "trokuta" u grafu,...) i neka je β_i skup statističkih parametara. Prema [7], **model eksponencijalnih slučajnih grafova** je skup svih mogućih grafova s n vrhova u kojem se svaki graf G pojavi s vjerojatnošću

$$P(G) = \frac{1}{Z} \exp\left(-\sum_i \beta_i \epsilon_i\right) \tag{3.5}$$

gdje je

$$Z = \sum_G \exp\left(-\sum_i \beta_i \epsilon_i\right) \tag{3.6}$$

Prema [3], statistički paramteri β_i određuju se tako da vrijednosti mjerljivih svojstava iz stvarne mreže, ϵ_i^*, budu jednake odgovarajućoj prosječnoj vrijednosti po skupu grafova,

$$\langle \epsilon_i \rangle = \sum_G \epsilon_i(G) P(G) = \epsilon_i^* \tag{3.7}$$

Nakon što je vjerojatnost $P(G)$ određena, možemo ju koristiti kako bismo izračunali neka druga mjerljiva svojstva. Očekivana vrijednost svojstva x dana je s

$$\langle x \rangle = \sum_G x(G) P(G) = \frac{1}{Z} \sum_G x(G) \exp\left(- \sum_i \beta_i \epsilon_i \right) \tag{3.8}$$

Ako je mjerljivo svojstvo x zapravo neko svojstvo ϵ_i iz skupa mjerljivih svojstva za skup grafova, prosječna vrijednost tog svojstva dana je s

$$\langle \epsilon_i \rangle = \sum_G \epsilon_i(G) P(G) = \frac{1}{Z} \sum_G \epsilon_i \exp\left(- \sum_i \beta_i \epsilon_i \right) = \frac{\partial F}{\partial \beta_i} \tag{3.9}$$

gdje je

$$F = \ln Z \tag{3.10}$$

Poglavlje 4

Analiza i modeliranje mreže suradnje

U sljedećem poglavlju iznijet ćemo podatke koji su korišteni za izradu jedne socijalne mreže - mreže znanstvene suradnje, softver koji je korišten za izradu takve mreže te analizu i modeliranje mreže.

4.1 Programsko rješenje

Za kompletno programsko rješenje korišten je R[1] i paketi u R-u. Za web scraping sa servisa Google Scholar korišteni su paketi scholar[2] i rvest[3]. Za izradu mreža, analizu i vizualizacije mreža korišten je paket igraph[4]. Za prikaz distribucija i histograma korišten je paket ggplot2[5].

4.2 Podaci

Google Scholar[6] je Google-ov servis za pretraživanje znanstvenih radova. Na servisu su dostupni znanstveni časopisi, sažeci, recenzirani članci, magistarski i doktorski radovi, knjige i još mnogo toga. Za svaki znanstveni rad Google Scholar dohvaća bibliografske podatke tako što automatizirano pretražuje i indeksira sadržaj sa Web-a (engl. web crawling). Na servis Google Scholar nadograđen je servis Google Scholar Citation[7]. Na tom servisu znanstvenicima se omogućuje da kreiraju osobne profile sa svojim radovima

[1] http://www.r-project.org/
[2] https://cran.r-project.org/web/packages/scholar/scholar.pdf
[3] https://cran.r-project.org/web/packages/rvest/rvest.pdf
[4] https://cran.r-project.org/web/packages/igraph/igraph.pdf
[5] https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf
[6] https://scholar.google.com/
[7] https://scholar.google.com/citations

i brojem citiranja za pojedini rad, ručno ili automatski. Na taj način smanjuju se greške pri pretraživanju i indeksiranju sadržaja od strane Google Scholar-a. Svaki znanstvenik na svom profilu može navesti znanstvene discipline, instituciju, sveučilište (ako pripada nekom) i koautore. Zato na servisu Google Scholar Citation postoje i stranice posvećene disciplinama i one posvećene sveučilištima.
Google Scholar omogućava pretraživanje znanstvenih radova ili pretraživanje profila znanstvenika. Radovi se mogu pretraživati po sljedećim kriterijima[8]:

1. ako sadržavaju riječi
2. ako sadržavaju točan izraz
3. ako sadržavaju barem jednu riječ
4. ako ne sadržavaju neke riječi
5. pojavljuju li se prethodne riječi bilo gdje u radu ili u naslovu rada
6. autori rada
7. u čemu je rad objavljen
8. koji je raspon godina

Profili znanstvenika mogu se, uz pretragu po riječima, pretraživati po disciplini i po sveučilištu.[9] Pri pretraživanju po riječima, podudaranje se traži u imenu ili pripadnosti instituciji.

Podaci za izradu kompleksne mreže sastoje se od completeAuthors.txt tekstualne datoteke s 73 053 linija, gdje svaka linija predstavlja popis autora za neki znanstveni rad. Tekstualna datoteka dobivena je web scraping-om, tehnikom prikupljanja podataka s web stranica. Odabrane su tri discipline, ”robotics”, ”Internet of things” i ”computer science, što je u prijevodu robotika, internet stvari i računalna znanost. Za svaku disciplinu preuzeto je po 500 najcitiranijih znanstvenika, te nakon toga do 200 najcitiranijih znanstvenih radova po znanstveniku. Od tih radova, uzeti su samo radovi napisani između 2010. i 2015. godine. To znači da skup podataka ne sadrži sve radove za sve znanstvenike i također ne sadrži sve

[8] izgled URL-a za pretraživanje s redom navedenim kriterijima:
`https://scholar.google.com/scholar?as_q=PRVI&as_epq=DRUGI&as_oq=TRECI&as_eq=CETVRTI&as_occt=PETI&as_sauthors=SESTI&as_publication=SEDMI&as_ylo=OSMI_OD&as_yhi=OSMI_DO` (peti kriterij može biti ”any” za opciju bilo gdje u radu, odnosno ”title” za opciju u naslovu)

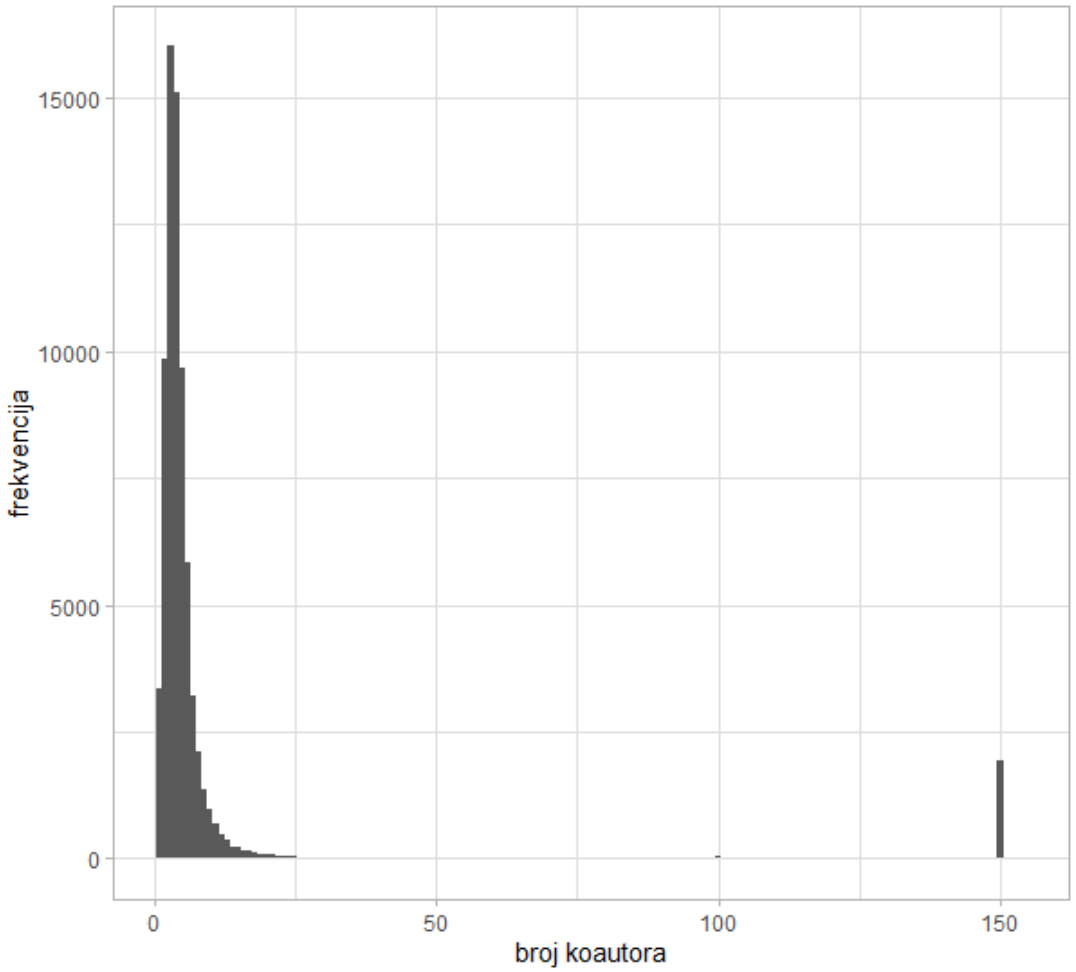
[9] izgled URL-a za pretraživanje su
`https://scholar.google.com/citations?&view_op=search_authors&mauthors=RIJECI`,
`https://scholar.google.com/citations?&view_op=search_authors&mauthors=label:robotics+RIJECI` ili `https://scholar.google.com/citations?view_op=view_org&org=ID`.
ID je identifikacijski broj sveučilišta.

Radovi	broj	postotak
– s jednim autorom	3362	4.60 %
– s dva autora	9842	13.47 %
– s tri autora	16 020	21.93 %
– s četiri autora	15 121	20.70 %
– s pet autora	9666	13.23 %
– ostalo	19 042	26.07 %
Ukupno radova	73 053	100.00 %
Ukupno autora	112 846	
Prosječno radova po autoru	5.79	
Prosječno koautora po radu	8.92	
Najveći broj koautora u radu	145	

Tablica 4.1: Analiza skupa podataka prije izgradnje mreže

znanstvenike u određenoj disciplini već samo one najutjecajnije. Takav skup podataka dat će nam uvid u suradnju najutjecajnijih znanstvenika na njihovim najutjecajnijim radovima i pokazat će nam jesu li oni međusobno povezani ili povezani preko nekih drugih znanstvenika s kojima su surađivali. Prema [2], veći broj suradnika u znanstvenom radu vodi do većeg broja citiranja tog rada, što znači da su radovi koje smo isključili bili samostalni ili s vrlo malim brojem suradnika (koautora) pa neće značajno utjecati na bitna svojstva mreže. Svaki rad sadrži popis autora u obliku punih imena i prezimena znanstvenika ili inicijala imena i punog prezimena, gdje su različiti autori odvojeni zarezom. Iz tog razloga, svaki popis autora pretvorimo na način da su za svakog autora navedeni inicijali imena i puno prezime. Treba naglasiti da su u takvom pristupu moguće greške. Ukoliko neki autor ima više od jednog imena, može se dogoditi da je u nekom popisu autora za znanstveni rad naveden sa svim imenima, dok u nekom drugom popisu samo s jednim. Opisani pristup će u takvom slučaju razlikovati dva autora, a zapravo je to ista osoba. Također, moguće je da neki popis autora sadrži i ime institucije s koje je znanstvenik, odvojeno zarezom. U tom slučaju, institucija će se uzeti kao jedan od autora.

Sa histograma na slici 4.2 zaključujemo da najveći broj znanstvenih radova ima tri (ko)autora, njih 16 020 što čini 21.93% skupa podataka. Nešto manje ima radova s četiri (ko)autora, 15 121 što čini 20.70%. Unatoč tome što smo skupljanjem podataka potencijalno isključili radove s vrlo malim brojem koautora, vidimo na 4.2 da je njihova frekvencija svejedno najveća u podacima. Iz tablice 4.1 i histograma na slici 4.2 vidimo da je najveći broj (ko)autora nekog znanstvenog rada jednak 150 i tih radova ima 1936. Ti radovi napisani su od istih nekoliko skupina autora, gdje su teme radova unutar svake skupine autora vrlo slične ili nastavci na prošli rad. Analizom autora po broju radova dobivena



Slika 4.1: Histogram frekvencija za broj koautora

je tablica 4.2 iz koje vidimo da je znanstvenik koji se pojavljuje u najvećem broju radova Z1, koji se pojavljuje u čak 1022 radova, što čini 1.40% svih radova.

Iz svakog popisa autora izvučeni su svi mogući parovi autora. Nakon toga, stvoren je neusmjereni graf bez težina s 112 456 vrhova i 24 072 350 bridova. Vrhovi u grafu predstavljaju autore koji su povezani bridom samo ako su zajedno navedeni kao (ko)autori barem jednog znanstvenog rada. Ukoliko je neki popis autora imao samo jednog navedenog autora, ne postoje parovi autora te takvi popisi nisu uzeti u obzir pri stvaranju mreže.

znanstvenik	broj	postotak
Z1	1022	1.40 %
Z2	702	0.96 %
Z3	685	0.94 %
Z4	681	0.93 %
Z5	681	0.93 %

Tablica 4.2: 5 znanstvenika koji se pojavljuju u najvećem broju radova

4.3 Analiza mreže

broj vrhova (n)	112 456
broj bridova (m)	24 072 350
prosječni najkraći put (ℓ)	4.01
promjer (d_G)	11
prosječni koeficijent klasteriranja (C)	0.85
gigantska kompomenta - broj vrhova	108 873
srednji stupanj $\langle k \rangle$	428.1203

Tablica 4.3: Osnovne veličine mreže

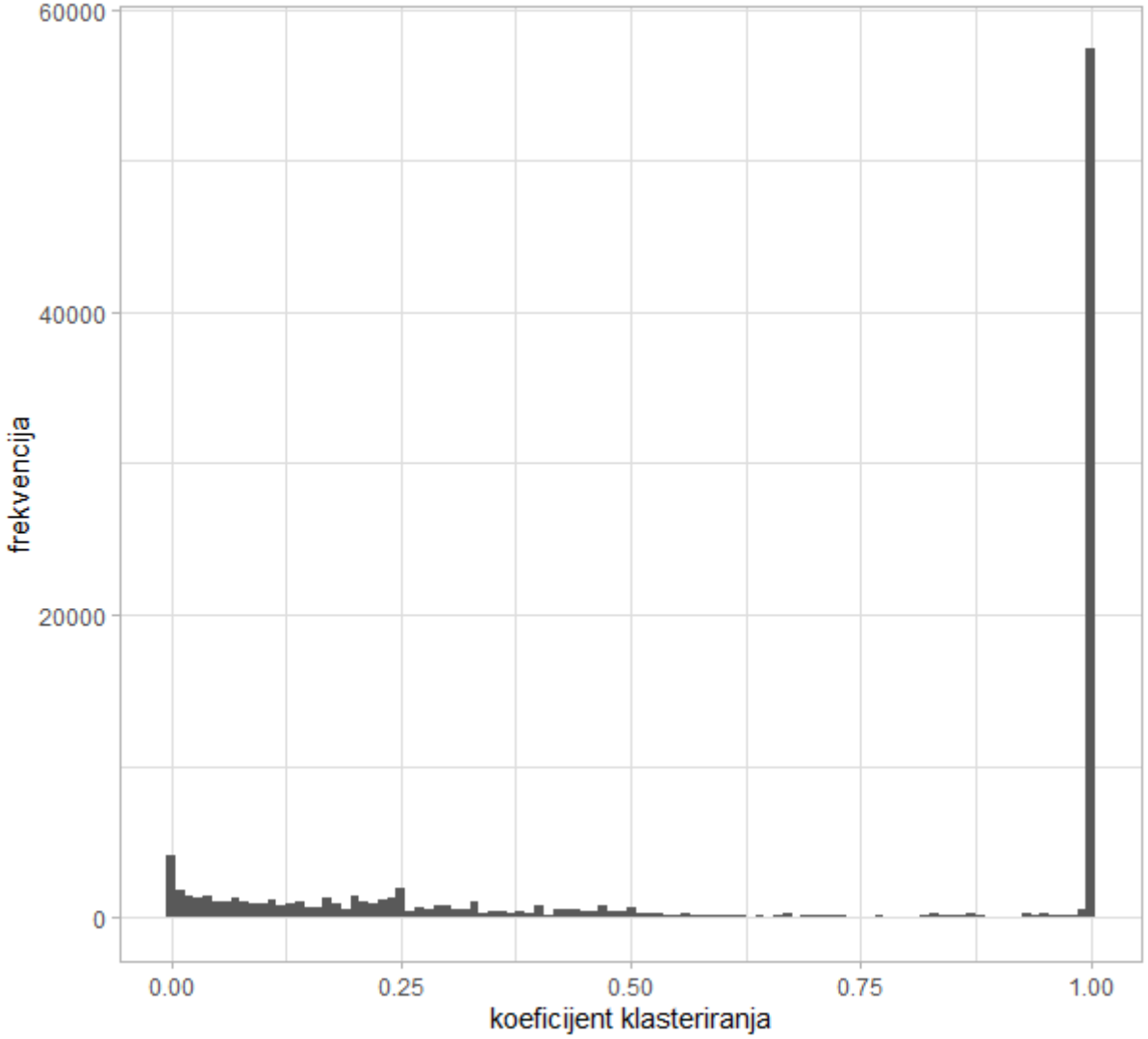
Svojstvo malog svijeta

Ova mreža nije povezana, što znači da postoji bar jedan par vrhova između kojih ne postoji brid. Treba napomenuti da pri računanju prosječne duljine najkraćeg puta i promjera iz računa izostavljamo vrhove između kojih ne postoji put. Prosječna duljina najkraćeg puta ℓ iznosi 4.01 što je manje od iznosa iz istraživanja navedenog u 2.1 i manje od $\log_2 n$. No da bismo zaključili ima li ova mreža svojstvo malog svijeta, trebali bismo znati kako se veličina ℓ ponaša s obzirom na različite veličine mreže, to jest različite n (definicija 2.1.2), a to ne možemo jer imamo samo jednu fiksnu veličinu mreže. Ono što možemo zaključiti jest da su vrijednosti prosječne duljine najkraćeg puta i promjera vrlo male s obzirom na n.

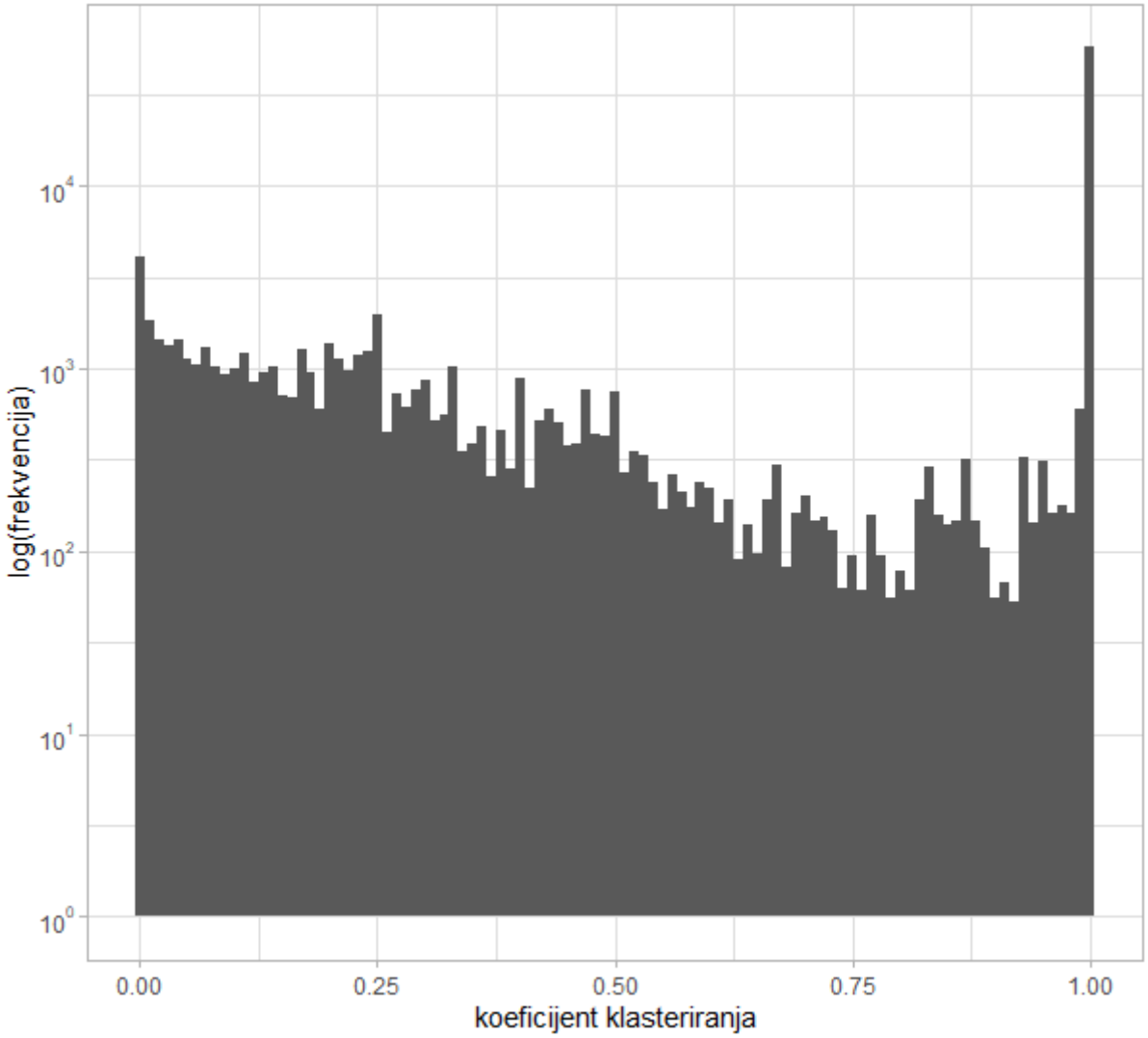
Koeficijent klasteriranja

Prosječni koeficijent klasteriranja vrlo je visok, što je logično s obzirom na veliki broj vrhova s najvišom vrijednosti koeficijenta klasteriranja, njih čak 57 350 (51%). Razlog tome

je veliki broj znanstvenih radova s tri ili više autora koji sudjeluju u pisanju te takvi radovi stvaraju ”trokute” u mreži. Prema [6], objašnjenje je sljedeća situacija - znanstvenici si međusobno predstavljaju svoje suradnike (koautore) ili ih spaja institucija. Na slikama 4.2 i 4.3 možemo vidjeti dvije verzije histograma za koeficijent klasteriranja. Na drugoj slici histogram je napravljen s logaritmiranim frekvencijama kako bi se bolje vidjeli stupci, te će se takav histogram koristiti u analizi ostalih veličina gdje bude potrebno.



Slika 4.2: Histogram koeficijenata klasteriranja (1)

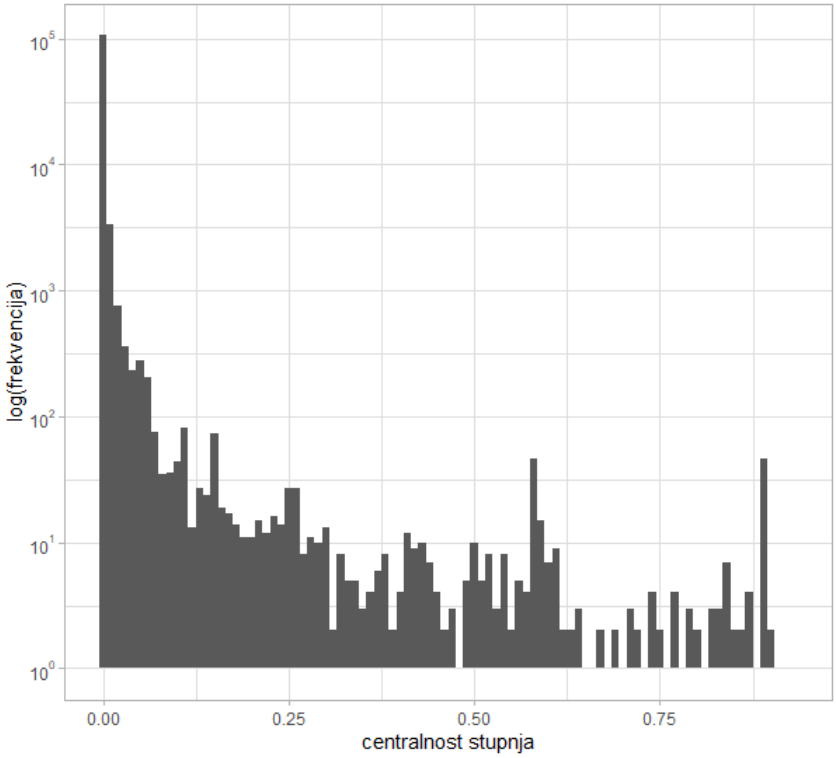


Slika 4.3: Histogram koeficijenata klasteriranja (2)

Centralnosti

znanstvenik	vrijednost
Z2	0.9301
Z5	0.9011
Z4	0.9011
Z6	0.8922
Z7	0.8920

Tablica 4.4: 5 znanstvenika najveće centralnosti stupnja



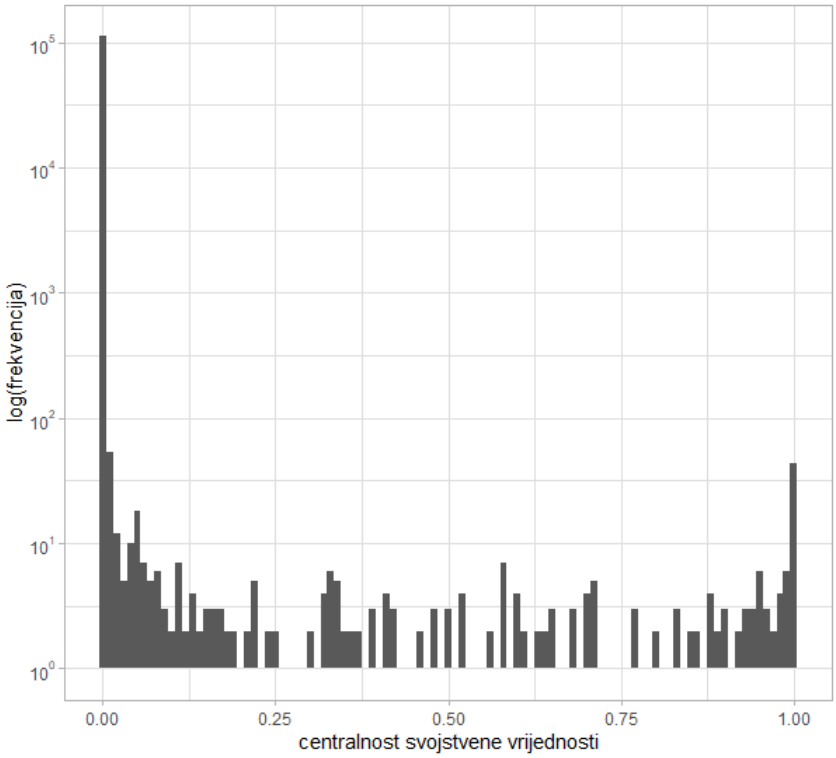
Slika 4.4: Histogram centralnosti stupnja

Velika većina vrhova u ovoj mreži ima male vrijednosti centralnosti stupnja, što znači da prema toj centralnosti postoji manjina vrhova koja je vrlo bitna u mreži. U tablici 4.4 nalazi se pet najbitnijih znanstvenika s obzirom na centralnost stupnja. Možemo primijetiti da je

najbitniji je Z2 te da se trojica iz te tablice nalaze i u tablici 4.2.

znanstvenik	vrijednost
Z5	1.0000000
Z6	0.9998397
Z8	0.9998396
Z9	0.9998396
Z3	0.9998396

Tablica 4.5: 5 znanstvenika najveće centralnosti svojstvene vrijednosti

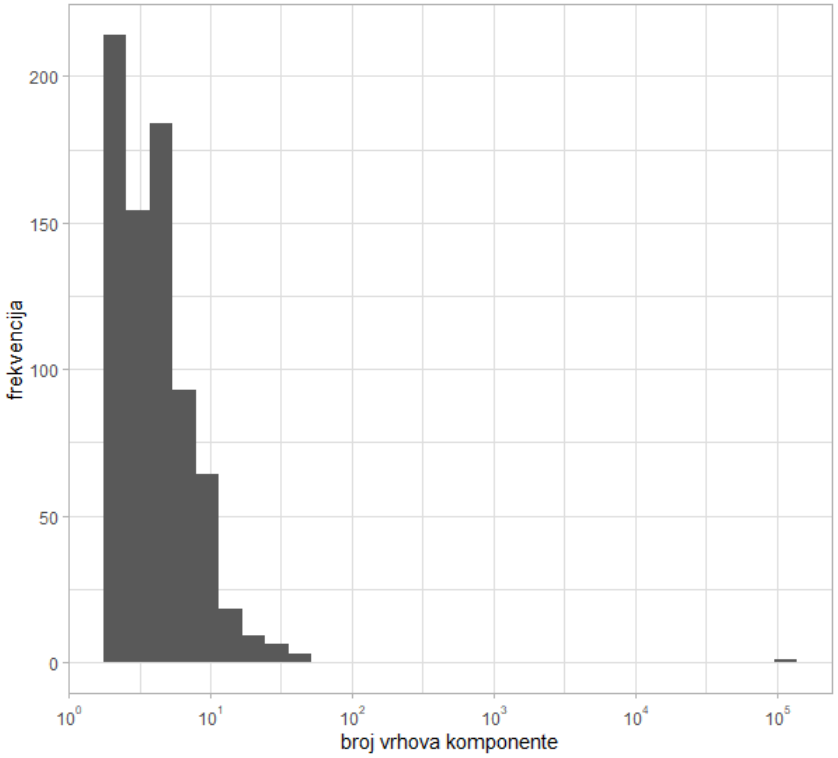


Slika 4.5: Histogram centralnosti svojstvene vrijednosti

Od znanstvenika iz tablice 4.5, dva znanstvenika pojavljuju se i u tablici 4.2. Uspoređujući tablicu 4.5 s tablicom 4.4, vidimo da se i tu podudaraju samo dva znanstvenika, pa možemo zaključiti da su Z5 i Z6 među najcentralnijim znanstvenicima po broju suradnji i po suradnji s drugim znanstvenicima koji su centralni.

Pri analizi izračunatih vrijednosti centralnosti blizine, otkriveno je da postoji 3583 vrijednosti koje se nalaze između 8.892×10^{-6} i 8.897×10^{-6} te 108 873 vrijednosti između 2.784×10^{-4} i 2.787×10^{-4}. To znači da u mreži postoji manjina vrhova koji imaju jako malu duljinu najkraćeg puta do ostalih vrhova u mreži i većina koja ima nešto veću. Rezultat je takav s obzirom na malenu vrijednost prosječne duljine najkraćeg puta i što ni ovdje nismo uzeli u obzir vrhove između kojih ne postoji put. Pošto ima 428 vrhova s najmanjom vrijednosti centralnosti blizine u ovom grafu, rangiranje nema smisla pa nećemo analizirati pet najmanjih vrijednosti.

Gigantska komponenta



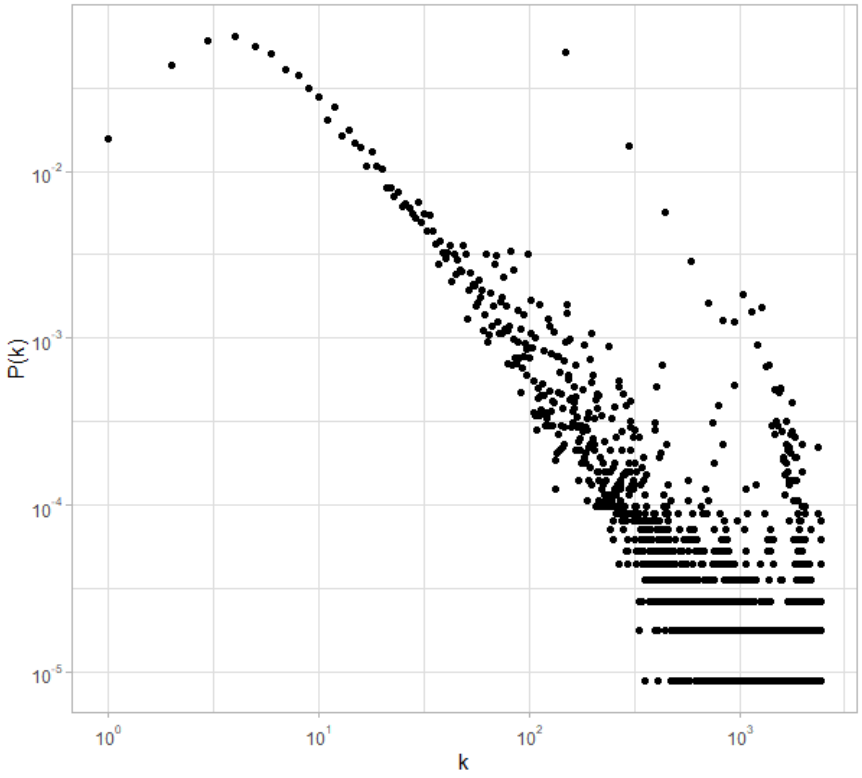
Slika 4.6: Histogram veličina komponenti povezanosti

Ova mreža sastoji se od 746 komponenti povezanosti, od kojih najveća ima 118 873 vrhova, dok ostale 745 komponente imaju od 1 do 50 vrhova, od kojih najviše ima komponenti s 2 vrha (njih 214). Na slici 4.6 možemo vidjeti histogram veličina komponenti. Vrhovi iz najveće komponente su isti oni vrhovi kojima je centralnost blizine između 2.784×10^{-4} i

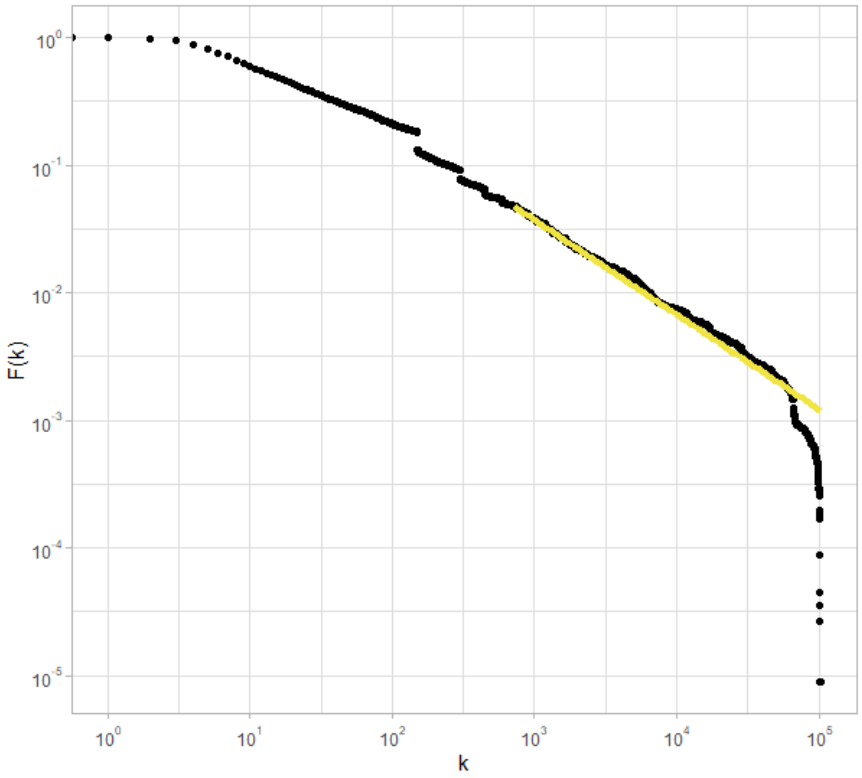
2.787×10^{-4}. Vidimo da je veličina najveće komponente mnogo veća od slijedeće najveće komponente(razlika između njih je 108 823 te da postoji samo jedna toliko velika komponenta, pa možemo zaključiti da je to gigantska komponenta ove mreže.

Distribucija stupnja

Na slici 4.7 možemo vidjeti distribuciju stupnja za ovu mrežu u logaritamskoj skali. Zaključujemo da ima "težak rep" i da također postoji manjina točki koje iskaču od ostatka te prividno stvaraju neku drugu distribuciju. Ova distribucija daje dojam da ovdje postoje dvije distribucije koje su nalik dvije različite distribucije zakona potencija. U R-u dobiven je parametar $\alpha = 1.742229$ za aproksimaciju zakonom potencija. Na slici 4.8 nalazi se kumulativna distribucija stupnja i aproksimacija dobivenim zakonom potencija.



Slika 4.7: Distribucija stupnja



Slika 4.8: Kumulativna distribucija stupnja i aproksimacija distribucijom zakona potencija

4.4 Modeliranje mreže

U ovom ćemo poglavlju pokušati aproksimirati mrežu s nekoliko ranije spomenutih modela mreža. U R-u su generirani modeli mreža, izračunata su svojstva, zatim su t-testom uspoređene distribucije za centralnost stupnja i centralnost svojstvene vrijednosti. Neće se uspoređivati vijednosti za centralnosti blizine jer su grafovi za modele povezani pa postoji velika razlika u vrijednostima na temelju koje možemo zaključiti da nisu iz iste distribucije.

Model Poissonovog slučajnog grafa

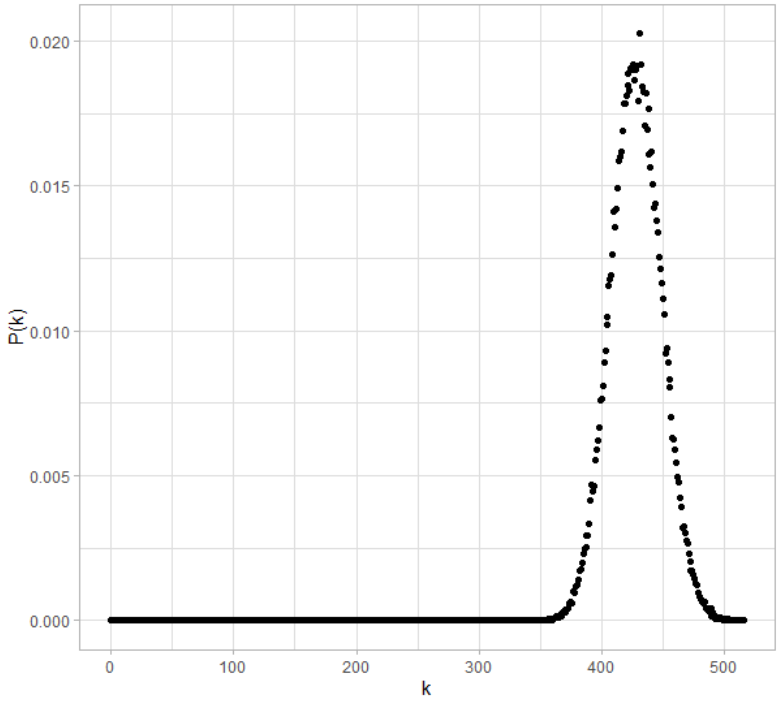
broj vrhova (n)	112 456
broj bridova (m)	24 068 245
prosječna duljina najkraćeg puta (ℓ)	2.19
promjer (d_G)	3
prosječni koeficijent klasteriranja (C)	0.0038
gigantska kompomenta - broj vrhova	/

Tablica 4.6: Osnovne veličine mreže za model Poissonovog slučajnog grafa

Ovaj graf je generiran tako što je izračunat $p = 0.003807003$ pomoću (3.2). Dobiveni graf je povezan i ima 24 068 245 bridova, što nije puno manje od mreže suradnje. Prosječna duljina najkraćeg puta i promjer manji su od onih u mreži suradnje, dok je prosječni koeficijent klasteriranja vrlo malen pa ne aproksimira dobro vrijednost iz mreže suradnje, što je bilo očekivano. Kao što je već spomenuto, distribucija u ovom modelu je Poissonova, pa ne aproksimira dobro distribuciju mreže suradnje. U tablici 4.7 možemo vidjeti p-vrijednosti t-testova za distribucije centralnosti i možemo zaključiti da ovaj model dobro aproksimira samo centralnosti stupnja.

t-test jednakosti očekivanja	mreža	model	p-vrijednost
centralnost stupnja	0.003807037	0.003806388	0.9952
centralnost svojstvene vrijednosti	0.001338364	0.790896308	2.2×10^{-16}

Tablica 4.7: Test jednakosti očekivanja za model Poissonovog slučajnog grafa



Slika 4.9: Distribucija stupnja za model Poissonovog slučajnog grafa

Konfiguracijski model

broj vrhova (n)	112 456
broj bridova (m)	8 741 367
prosječna duljina najkraćeg puta (ℓ)	2.60
promjer (d_G)	5
prosječni koeficijent klasteriranja (C)	0.70266
gigantska kompomenta - broj vrhova	/

Tablica 4.8: Osnovne veličine mreže za konfiguracijski model

Ovaj graf je generiran pomoću distribucije stupnja za mrežu suradnje, pa sigurno dobro aproksimira distribuciju stupnjeva za model. Primijetimo da je broj bridova više od triput veći i da je graf povezan. Prosječna duljina najkraćeg puta i promjer su vrlo mali. Prosječni koeficijent klasteriranja je nešto manji od koeficijenta u mreži, ali opet velik. Iz tablice 4.8 možemo zaključiti da model dobro ne aproksimira centralnosti.

t-test jednakosti očekivanja	mreža	model	p-vrijednost
centralnost stupnja	0.003807037	0.001382445	2.2×10^{-16}
centralnost svojstvene vrijednosti	0.001338364	0.029017505	2.2×10^{-16}

Tablica 4.9: Test jednakosti očekivanja za konfiguracijski model

Model malog svijeta

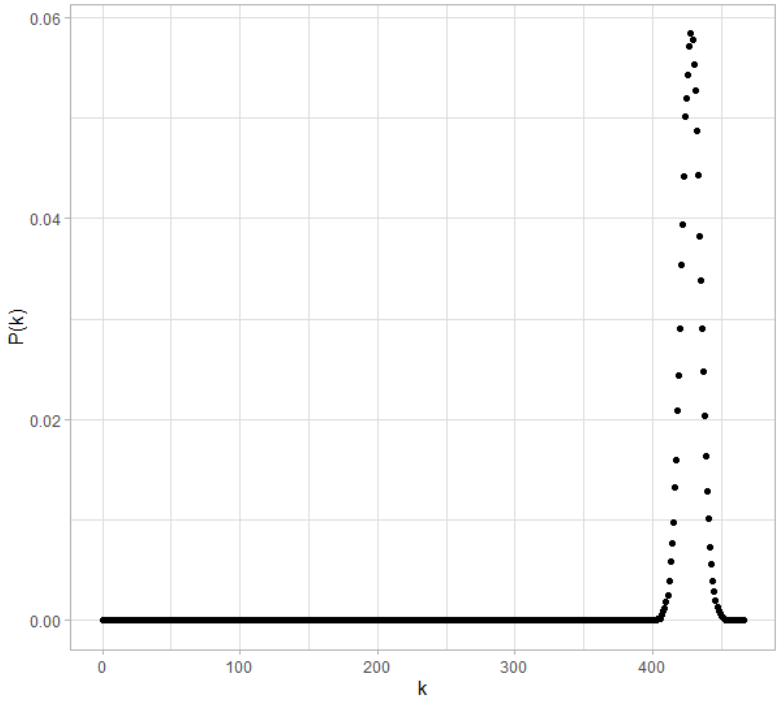
broj vrhova (n)	112 456
broj bridova (m)	24 065 584
prosječna duljina najkraćeg puta (ℓ)	2.70
promjer (d_G)	3
prosječni koeficijent klasteriranja (C)	0.5238106
gigantska kompomenta - broj vrhova	/

Tablica 4.10: Osnovne veličine mreže za model malog svijeta

Za ovaj model izračunat je parametar $m = 214.06015$ pomoću $\langle k \rangle = 2m$, a parametar $p = 0.0579$ određen je isprobavanjem njegovih raznih vrijednosti dok nije uspješno aproksimiran prosječni koeficijent klasteriranja. Prosječna duljina najkraćeg puta i promjer su i u ovom modelu manji nego u mreži. Na slici 4.10 vidimo distribuciju stupnja za model malog svijeta. Distribucija je Poissonova kao i kod modela Poissonovog slučajnog grafa.

t-test jednakosti očekivanja	mreža	model	p-vrijednost
centralnost stupnja	0.003807037	0.003805967	0.9921
centralnost svojstvene vrijednosti	0.001338364	0.924752254	2.2×10^{-16}

Tablica 4.11: Test jednakosti očekivanja za model malog svijeta



Slika 4.10: Distribucija stupnja za model malog svijeta

Poglavlje 5

Zaključak

Sa servisa Google Scholar preuzeto je do 200 najcitiranijih znanstvenih radova za svakog od 500 najcitiranijih autora iz područja robotike, Interneta stvari i računalne znanosti od 2010. do 2015. godine, što je rezultiralo popisom autora za 73 053 znanstvena rada. Taj skup podataka detaljno je analiziran. Otkriveno je da najveći broj radova ima tri koautora i najveći broj koautora nekog rada iznosi 145. Najčešći autor među radovima je Z1. Nakon analize, od tih podataka stvorena je mreža suradnje tj. koautorstva.

Mreža ima 112 456 vrhova i 24 072 350 bridova. Nakon toga izračunata su neka najbitnije veličine te su proučena svojstva. S obzirom na svoju veličinu, mreža ima jako malu prosječnu duljinu najkraćeg puta i promjer. Iako intuitivno vidimo da bi ova mreža trebala posjedovati svojstvo malog svijeta, nismo mogli matematički dokazati. Koeficijent klasteriranja je očekivano jako velik jer si znanstvenici međusobno predstavljaju druge znanstvenike ili se poznaju s iste institucije i tako stvaraju "trokute" u mreži. Što se tiče centralnosti stupnja, većina vrhova u mrežu ima malu centralnost, a najveću centralnost ima Z2. Neki vrhovi koji imaju centralnost stupnja koja je među najvećima, također se nalaze u popisu znanstvenika koji se nalaze u najviše znanstvenih radova. Za centralnost svojstvene vrijednosti zaključujemo da su Z5 i Z6, osim što su centralni po broju suradnje s mnogo različitih znanstvenika (centralni u smislu centralnosti stupnja), također su centralni po tome što su im susjedi centralni u tom smislu. Analiza centralnosti blizine daje nam vrlo male vrijednosti koje variraju oko 8.894×10^{-6} ili 2.786×10^{-4}. Analizom komponenti grafa, otkriveno je da vrijednosti za ogromnu komponentu u kojoj se nalazi 108 873 variraju oko 2.786×10^{-4}. Zaključujemo da znantvenici u mreži jako malo udaljeni jedni od drugih u smislu centralnosti blizine i približno su jednako mnogo su centralni. Proučena je i distribucija stupnja te je pokušana aproksimacija distribucijom zakona potencija koja je najčešća distribucija u stvarnom svijetu. Ta distribucija ne može u potpunosti opisati cijelu distribuciju stupnja mreže te također ima koeficijent koji nije među najčešćima u mrežama

stvarnog svijeta.

Svaki od odabranih modela može aproksimirati samo neki podskup veličina i svojstava mreže. Model Poissonovih slučajnih grafova i model malog svijeta dobro aproksimiraju broj bridova te centralnost stupnja, dok konfiguracijski model dobro aproksimira prosječni koeficijent klasteriranja. Naravno, ne postoji model koji bi u potpunosti dobro opisivaju bilo koju mrežu iz stvarnog svijeta.

Ova mreža vrlo je specifična zbog načina prikupljanja podataka pa zato ne daje sve uobičajene rezultate kao i ostale proučavane mreže suradnje među znanstvenicima. Nisu prikupljeni svi znanstveni radovi u bazi, disciplini ili za znastvenika, pa je rezultat ovakva mreža s mnogo komponenti povezanosti. U ovom radu obrađene su samo osnove teorije kompleksnih mreža i izvedeni su neki osnovni zaključci, stoga ovaj rad ostavlja prostora za buduće upotpunjavanje skupa i mreže koja prikazuje neku kompletniju strukturu sustava znantvenika na servisu Google Scholar.

Bibliografija

[1] UC Davis DataLab, *Creating Co-Author Networks in R*, https://datalab.ucdavis.edu/2019/08/27/creating-co-author-networks-in-r/.

[2] W. Figg, L. Dunn, D. Liewehr, S. Steinberg, P. Thurman, C. Barrett i J. Birkinshaw, *Scientific Collaboration Results in Higher Citation Rates of Published Articles*, Pharmacotherapy **26** (2006), 759–67.

[3] A. Fronczak, *Exponential Random Graph Models*, Encyclopedia of Social Network Analysis and Mining (2018), 810–826, http://dx.doi.org/10.1007/978-1-4939-7131-2_233.

[4] G. Ghoshal, *Structural and Dynamical Properties of Complex Networks.*, Disertacija, siječanj 2009, http://hdl.handle.net/2027.42/64757.

[5] Z. Huang i B. Yuan, *Mining Google Scholar Citations: An Exploratory Study*, (2012), 182–189.

[6] M. E. J. Newman, *The structure of scientific collaboration networks*, Proceedings of the National Academy of Sciences **98** (2001), br. 2, 404–409, ISSN 0027-8424, https://www.pnas.org/content/98/2/404.

[7] ———, *The structure and function of complex networks*, SIAM review **45** (2003), br. 2, 167–256, https://epubs.siam.org/doi/pdf/10.1137/S003614450342480.

[8] H. Petric Maretić, *Dinamička analiza mreža – evolucija istraživačkog područja temeljem radova objavljenih u sklopu DESIGN konferencije 2002–2014*, (2014), https://apps.unizg.hr/rektorova-nagrada/javno/stari-radovi/2760/preuzmi.

[9] ———, *Dinamički procesi na kompleksnim mrežama (Diplomski rad)*, (2014), https://urn.nsk.hr/urn:nbn:hr:217:651545.

[10] D. Strmečki, *Analiza točnosti pretraživanja (Diplomski rad)*, (2020).

Sažetak

U ovom radu analiziramo jednu mrežu suradnje među znanstvenicima. Podaci su preuzeti su sa servisa Google Scholar, te su očišćeni i analizirani. Iz tih podataka stvorena je mreža na način da su dva znanstvenika povezana ako su barem jednom obojica navedeni kao autori nekog znanstvenog rada. Za tu mrežu računamo određene veličine kako bismo proučili svojstva malog svijeta, tranzitivnosti, centralnosti, posjedovanje gigantske komponente i distribuciju stupnja. Mreža pokazuje visoku vrijednost za svojstvo tranzitivnosti, posjeduje gigantsku komponentu i distribucija stupnja ne prati u potpunosti neku poznatu distribuciju. Ne možemo dokazati da mreža posjeduje svojstvo malog svijeta, no mreža ima vrlo mali promjer i prosječnu duljinu najkraćeg puta. Proučavamo vrijednosti za tipove centralnosti te uspoređujemo pet najvećih vrijednosti. Nakon toga aproksimiramo mrežu s par teorijskih modela kao što su model Poissonovog slučajnog grafa, konfiguracijski model i model malog svijeta.

Summary

In this thesis we analyse a scientific collaboration network. The data is downloaded from the service Google Scholar, cleaned and then analysed. The network is created from that data so that two scientists are connected if they are at least once listed as the authors of a paper. For that network we calculate certain quantities so we can study the small-world property, transitivity, centrality measures, existence of a giant component and degree distribution. The network shows high value for transitivity, has a giant component and the degree distribution does not follow completely any known distribution. We can't prove that the network has the small-world property but nevertheless it has low values for the diameter and average shortest path length. We analyse values for different types of centrality and we compare five highest values. After that, we approximate the network with a few theoretical model such as Poisson random graph model, the configuration model and the small-world model.

Životopis

Rođena sam 17. veljače 1996. godine u Rijeci. Pohađala sam OŠ Maria Martinolića u Velom i Malom Lošinju te SŠ Ambroza Haračića, smjer opća gimnazija u Malom Lošinju. Akademske godine 2014./2015. upisujem preddiplomski studij Matematika na Prirodoslovno-matematičkom fakultetu Sveučilišta u Zagrebu, koji sam završila 2017. godine. U akademskoj godini 2017./2018. upisujem diplomski studij Računarstvo i matematika na istom fakultetu. Kao dio tima Outliers sudjelovala sam na natjecanju Mozgalo 2019. godine gdje smo osvojili šesto mjesto.