SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 2977

Usporedba algoritama otkrivanja zajednica u društvenim mrežama

Daniel Marić

Zagreb, svibanj 2022.

Umjesto ove stranice umetnite izvornik Vašeg rada.
Da bi ste uklonili ovu stranicu obrišite naredbu `\izvornik`.

Sadržaj

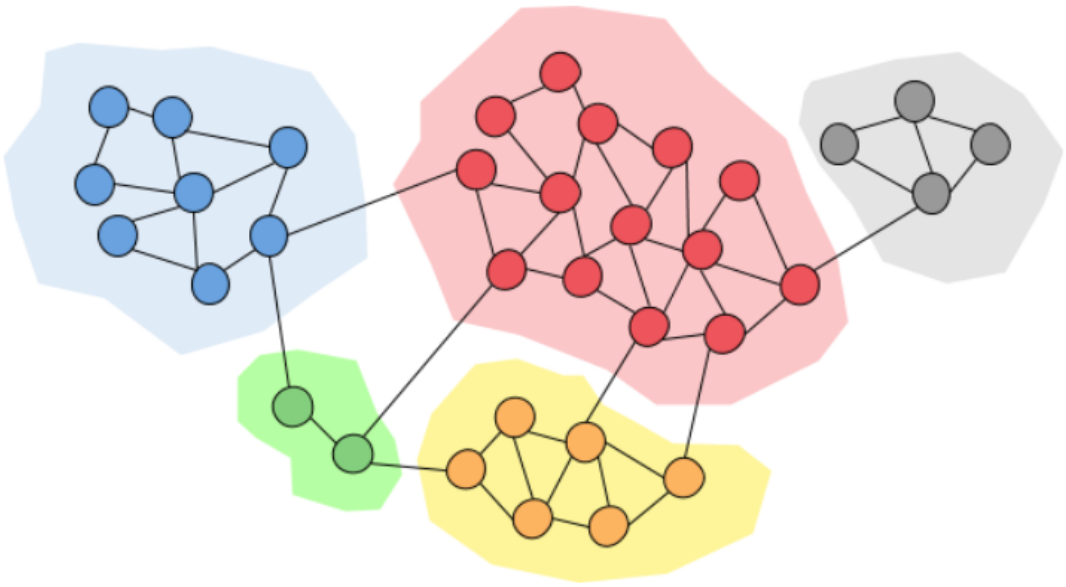
1. Uvod

Pojavom popularnih internetskih usluga za povezivanje korisnika stvorene su velike mreže društvenih zajednica. Generirane su velike količine podataka iz kojih je moguće izvući mnoštvo korisnih informacija. Takve zajednice sastoje se od puno manjih zajednica koje se po svojim karakteristikama razlikuju od ostalih. Takve zajednice potrebno je pronaći kako bi im se pristupilo na najbolji mogući način. Rješavanje ovog problema važno je i u drugim granama znanosti kao na primjer u sociologiji, biologiji ili računarskoj znanosti gdje su problemi predstavljeni na takav način, pomoću strukture grafa.

Upravo su grafovi najpogodnija struktura podataka za pristup ovome problemu gdje relevantne značajke, u primjeru društvenih mreža ljude, možemo prikazati pomoću čvorova dok će bridovi predstavljati veze između tih značajki. Više bridova među određenim značajkama značit će da tu mogu postojati obilježja zajednice, npr. u biologiji bi to mogla biti tkiva koja u organima obavljaju sličnu ulogu.

Rješavanje problema koji su predstavljeni grafovima je vrlo složeno, vremenski i prostorno. Ovakvi grafovi nisu jednostavnog oblika, ali u njima postoje određene pravilnosti koje se mogu iskoristiti. U tu svrhu razvijeno je mnogo algoritama za otkrivanje zajednica koji različitim pristupima pokušavaju pronaći rješenje ovog problema. Pojedini algoritmi su bolji od drugih na jednom tipu društvenih mreža ili lošiji na drugom te se zato koriste evaluacijske mjere kojima se procjenjuje koliko je dobro rješenje koje je algoritam pronašao. Što više algoritama se testira s različitim društvenim mrežama i evaluacijskim mjerama dobit ćemo bolji uvid u to kada je koji bolje koristiti. Najpoznatiji algoritam među njima je Newman-Girvanov algoritam koji će se nešto detaljnije opisati uz još nekoliko njih. Sličnim problemima bavili su se radovi Lancichinettija i Fortunata [2] iz 2016. te [3] iz 2009. godine.

Unatrag posljednih nekoliko godina uvedeni su zakoni o zaštiti osobnih podataka te je sada znatno teže dobiti pristup korisnim informacijama. Zato se



Slika 1.1: Primjer grafa nepreklapajućih društvenih zajednica.

koriste posebni algoritmi za generiranje umjetnih skupova podataka koji će za zadane parametre generirati graf pomoću kojih se mogu provoditi istraživanja.

U nastavku rada bit će opisana struktura i svojstva društvenih mreža i zajednica, algoritmi koji pronalaze društvene zajednice, skupovi podataka koji su korišteni u sklopu rada, programsko rješenje koje pokreće i evaluira rješenja algoritama te će se prikazati rezultati i do kojih se došlo.

2. Društvene mreže i zajednice

Društvene mreže moguće pronaći gdje god postoji sustav koji sadrži entitete koji su međusobno povezani. Primjera je mnogo, a neki od njih su: društvene web platforme, email mreže, web stranice koje sadrže poveznice prema drugima, uređaji koji su povezani preko internetske mreže i slično. Kako bi se skupina entiteta mogla nazvati društvenom zajednicom među njima mora postojati nekakav tip odnosa. Može biti jednosmjerni ili dvosmjerni te mogu postojati težine kojima se odnosu daje veća ili manja značajnost. Društvene mreže imaju složenu organizacijsku strukturu te se može pretpostaviti svojstvo lokalnosti koje kaže da ako jedan entitet ima veze prema neka druga dva entiteta onda je vjerojatnost da ta druga dva entiteta imaju vezu veća od prosječne.

Društvene mreže imaju karakteristično svojstvo grupiranja u strukturu zajednice. Ako se čvorovi mreže mogu podijeliti u nepreklapajuće ili preklapajuće zajednice tako da broj veza između članova zajednice značajno premašuje broj veza između bilo koje dvije zajednice znači da mreža ima strukturu društvenih zajednica. Mreže koje imaju takvu strukturu često se mogu prikazati i kao hijerarijske strukture. U ovom radu obradit će se mreže koje sadrže nepreklapajuće strukture sa vezama koje nemaju određene težine.

Proces pronalaska društvenih zajednica jedan je od glavnih zadataka u analizama društvenih mreža. Detekcija zajednica može biti vrlo korisna u raznim primjenama kao što je primjerice pronalaženje grupa kojima bi se mogle slati reklame za određene proizvode koji bi ih mogli zanimati umjesto da se svakom pojedincu šalju posebno. Još jedan primjer bio bi preporuka određenih sadržaja koji bi se mogli prikazivati grupama koje pokazuju zanimanja prema sličnim interesima. Primjera ima još mnogo, ali iz ova dva već je vidljivo da se korisne informacije mogu zaključivati iz društvenih mreža. Kako bi društvene mreže pohranili i analizirali u računalu potrebna je prikladna struktura podataka koja će u ovom slučaju biti graf.

2.1. Reprezentacija društvenih mreža

Graf je važna struktura podataka u području računarstva. Pomoću njega moguće je prikazati razne odnose i procese područja bioloških, društvenih i informacijskih sustava. Grafovima se modeliraju vrlo teški problemi kao primjerice problem kineskog poštara ili problem trgovačkog putnika koji je NP težak problem što znači da nema rješenje u polinomnom vremenu.

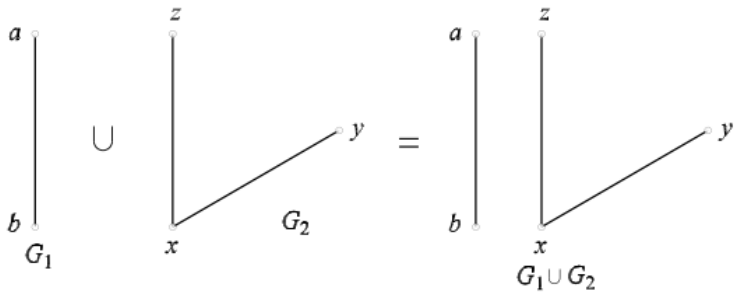
Prema definiciji jednostavan graf G sastoji se od nepraznog konačnog skupa *V(G)*, čije se elemente naziva vrhovi ili čvorovi grafa i konačnog skupa *E(G)* različitih dvočlanih podskupova skupa *V(G)* koji se naziva bridovima [5]. Graf može imati najviše $\frac{n(n-1)}{2}$. U radu će se razmatrati jednostavni grafovi koji nemaju petlje i više bridova između istih čvorova. Bridovi će biti bestežinski i neusmjereni.

Bitna definicija tiče se stupnja vrhova grafa. Stupanj vrha v grafa G je broj bridova koji su incidentni s v. Stupanj vrha označava se sa *deg(v)*. Vrh stupnja 0 zove se izolirani vrh, a vrh stupnja 1 krajnji vrh. [5]

Šetnja je graf sa skupom vrhova $V(G) = \{x_1, x_2, ..., x_l\}$ i bridova $E(G) = \{x_0x_1, x_1x_2, ..., x_{l-1}x_l\}$. vrhovi x_0 i x_l definiraju se kao krajevi dok je l duljina šetnje. Ako su svi bridovi šetnje različiti tada se ona naziva staza. Ako su uz to i svi vrhovi različiti onda se takva šetnju naziva putem. Ako put počinje i završava u istom vrhu tada graf sadrži ciklus. Uz pretpostavljena ograničenja najmanji ciklus koji graf u ovom radu može imati je trokut što je često obilježje društvenih mreža.

Definicija puta omogućava definiranje važnog koncepta koji će se pojavljivati u radu pojedinih algoritama. Ako u grafu za svaki par vrhova postoji barem jedan put koji ide od jednog do drugog onda je graf povezan. Ako između vrhova postoji više putova onda je najkraći onaj koji ima najmanju duljinu. Promjer ili dijametar povezanog grafa je najveća udaljenost između bilo koja dva vrha u grafu. Ako ipak postoji barem jedan par vrhova između kojih ne postoji put onda je graf podijeljen u barem dva podgrafa. Svaki maksimalno povezani podgraf zove se komponenta povezanosti. Primjer se može vidjeti na slici 2.1.

Grafovi se mogu pohranjivati u obliku matrice susjedstva gdje su dva vrha, i i j susjedna ako im je element matrice Aij jednak 1, a inače 0. Zbog pretpostavke da ne postoje petlje na dijagonali matrice susjedstva svi su elementi nule. Za reprezentaciju neusmjerenog grafa matrica susjedstva je simetrična što znači da je dovoljno pohraniti samo jedan trokut matrice, iznad ili ispod dijagonale. Suma



Slika 2.1: Primjer nepovezanog grafa

elemenata i-tog retka ili stupca jednaka je stupnju vrha i

Jednostavniji oblik pohrane lista susjedstva koja se koristi tako da se pohranjuje skup susjednih bridova koji predstavljaju graf. Lista susjedstva je prostorno učinkovitija od matrice susjedstva kada su u pitanju rijetki grafovi kod kojih većina vrhova nije međusobno povezana. Prostorno zauzeće ovisi o broju vrhova i bridova u grafu, dok je kod matrice susjedstva uvijek proporcionalno kvadratu broja vrhova.

2.2. Obilježja društvenih zajednica

Društvene zajednice moguće je definirati na nekoliko načina sa različitih stajališta, ali ne postoji niti jedna univerzalno prihvaćena definicija. Definiranje vrlo često ovisi o problemu koji se promatra zajedno sa specifičnim detaljima i primjenama gdje se pojam zajednice koristi. Prema radu [1] zajednice je moguće promatrati iz lokalne i globalne perspektive.

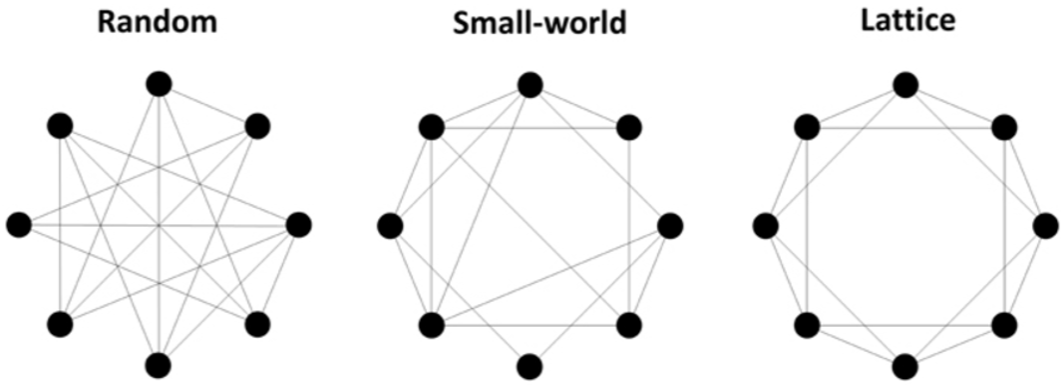
Iz lokalne perspektive zajednica se može promatrati kao grupa entiteta koji su međusobno sličniji u odnosu na ostale entitete skupa podataka. Zajednica se formira tako što slični elementi imaju mnogo više interakcija sa članovima unutar zajednice u odnosu na one izvan. Zajednica se može smatrati kao autonomna skupina te ima smisla u određenim situacijama evaluirati svaku zasebno od ostatka društvene mreže. Stroga definicija društvene mreže kaže kako je društvena zajednica podgraf u kojem su svi članovi međusobno u interakciji [4]. Takva definicija odgovara terminu klike u teoriji grafova koji označava skup vrhova koji su svi međusobno susjedni. Najjednostavniji primjer klike je trokut i oni se pojavljuju u svim društvenim mrežama. Veće klike od trokuta se pojavljuju rjeđe te ovakva definicija tako postaje manje praktična u stvarnim primjerima. Još jedan problem klike je to su tada svi vrhovi simetrični bez mogućnosti razlikovanja nji-

hovih svojstava. U praktičnim primjerima očekuje se da među vrhovima postoji određena hijerarhijska struktura sa više i manje važnim čvorovima. Moguće je relaksirati pojam klike. Mogućnost je iskoristiti doseg i duljinu puta između čvorova. n-klika je takav podgraf da niti jedan par vrhova nije međusobno udaljen za više od n koraka i skup je maksimalan u smislu da niti jedan drugi čvor nije udaljen za više od n od svakog čvora iz podgrafa. Može se primijetiti da članovi podgrafa mogu biti povezani preko posrednika koji nije član grupe te onda n-klika ipak nije dovoljno dobra definicija. Definicija n-klana to popravlja. n-klan je n-klika u kojoj je dijametar podgrafa manji ili jednak n. Takva definicija ima problem što u njoj i dalje postoji zahtjev n-klike te se tako dolazi do definicije n-kluba. n-klub je podgraf gdje je dijametar manji ili jednak n. Tada je i svaki n-klan i n-klub i n-klika.

Iz globalne perspektive zajednica se može definirati promatrajući graf u cjelini. Takve definicije koriste se u slučajevima kada su zajednice dijelovi sustava bez kojih bi njegovo funkcioniranje bilo značajno izmijenjeno. Definicije se najčešće izvode indirektno, iz algoritama prema kojem je neko svojstvo iskorišteno kako bi se zajednice otkrile. Moguće je definirati null model koji će odgovarati prema određenim strukturnim karakteristikama, ali inače je slučajni graf. Model se tada koristi za usporedbu kako bi se odredilo ima li promatrani graf strukture zajednica. Poznati null model grafa predložili su Newman i Girvan koji se dobije tako da se u početnom grafu slučajno prespajaju bridovi pod uvjetom da stupanj svakog vrha ostane isti kao u početnom grafu. Iz njega je proizašla definicija modularnosti, odnosno funkcije kojom je moguće ocijeniti kvalitetu pronađenih zajednica u grafu. Modularnost je važna mjera jer ima nekoliko primjena u području otkrivanja zajednica. Koristi se kao mjera koja određuje koliko su kvalitetno određene grupe u mreži, ali i kao sastavni dio poznatog Girvan-Newmanovog algoritma [1].

2.3. Small-world mreže

Small-world mreže imaju obilježja dva tipa mreža. Prva mreža je slučajna mreža za koju je karakteristično što je prosječna udaljenost između dva vrha vrlo mala. Druga mreža je rešetkasta u obliku prstena gdje je svaki čvor susjedan sa $\frac{n}{2}$ čvorova sa svake strane. Small-world mreža posjeduje svojstva tih grafova te se pomoću njih može procijeniti u kojoj je mjeri mreža zaista small-world. Na temelju tih svojstava nastao je i Watts–Strogatz model koji služi za generiranje



Slika 2.2: Primjeri slučajnog, small-world i rešetkastog grafa.

slučajnih grafova društvenih mreža što se može iskoristiti u testiranjima raznih algoritama za detekciju zajednica. Primjeri mreža prikazani su na slici 2.2.

Small-world mreža je graf u kojem većina čvorova nisu susjedi, ali susjedi nekog čvora imaju veliku vjerojatnost da su i oni susjedi te se do svakog čvora može doći kroz nekoliko koraka što znači da bilo koja dva čvora imaju kratku međusobnu udaljenost. Specifično je što se ona za dva slučajno izabrana čvora te za fiksiran prosječan stupanj vrha povećava proporcionalno logaritmu broj čvorova u grafu dok koeficijent grupiranja nije malen. Small-world mreže sadrže klike i grupe koje su gotovo klike što proizlazi iz visokog koeficijenta grupiranje. Društvene mreže posjeduju svojstva small-world mreže.

Koeficijent grupiranja je mjera stupnja u kojem čvorovi u grafu teže grupiranju. Postoje dvije verzije mjere, lokalna i globalna. U lokalnoj verziji mjera se računa za pojedini čvor te govori u kolikoj je on mjeri grupiran sa svojim susjedima. Mjera se za čvor i računa kao suma broja veza koje postoje između susjeda promatranog čvora podijeljeno sa brojem svih mogućih veza,

$$C_i = \frac{2 \mid e_{jk} : v_j, v_k \in N_i, e_{jk} \in E \mid}{k_i(k_i - 1)}. \tag{2.1}$$

Ako iz formule maknemo koeficijent 2 tada se ona može koristiti za usmjerene grafove. Globalni koeficijent grupiranja daje informaciju o grupiranju u cijeloj društvenoj mreži. Temelji se na trojkama čvorova. Trojku čine promatrani čvor i druga dva čvora. Ako su povezani sa dva brida zovu se otvorena trojka, a ako su povezani sa tri zovu se zatvorena trojka što znači da jedan trokut čine tri trojke. Koeficijent se tada računa kao broj zatvorenih trojki podijeljen sa ukupnim brojem trojki,

$$C = \frac{broj\ zatvorenih\ trojki}{ukupan\ broj\ trojki}. \tag{2.2}$$

Formula je primjenjiva i na usmjerene i neusmjerene grafove.

Kratka prosječna duljina puta između čvorova znači da postoje čvorovi sa velikim brojem veza odnosno visokim stupnjem. Takvi čvorovi nazivaju se sabirnice te služe kao posrednici u mnogim putevima između ostalih čvorova. Primjer iz stvarnog svijeta može se pronaći u zračnim letovima između gradova. Na putovanju između dva grada vrlo često nije potrebno više od tri leta jer mnogo letova ide preko jednog velikog grada sa puno letova prema drugima.

Koliko mreža pripada small-world mreži može se izraziti pomoću small-koeficijenta, σ, koji se računa tako da se uspoređuju koeficijent grupiranja i karakteristična duljina puta u mreži sa slučajnim grafom koji ima jednak prosječan stupanj vrhova. Za karakterističnu duljinu puta najčešće se koristi prosječna minimalna udaljenost između vrhova. Koeficijent se računa prema formuli:

$$\sigma = \frac{\frac{C}{C_r}}{\frac{L}{L_r}}. \tag{2.3}$$

C i L su mjera grupiranja i prosječna duljina puta u promatranoj mreži dok su C_r i L_r su mjera grupiranja i prosječna duljina puta u slučajnom grafu. Ako je $\sigma > 1$ tada se može smatrati da je mreža small-world. No mjera pokazuje lošu otpornost na rast broja čvorova u mreži [6].

Druga mjera kojom se može izmjeriti koliko je mreža small-world uspoređuje promatranu mrežu s mrežom rešetkastog oblika (eng. lattice network) i slučajnom mrežom. Mjera kombinira karakterističnu duljinu puta i koeficijent grupiranja sa koeficijentom grupiranja rešetkaste mreže i karakterističnom duljinom puta ekvivalentnog slučajnog grafa prema sljedećoj formuli:

$$\omega = \frac{L_r}{L} - \frac{C}{C_l} \tag{2.4}$$

Ovakva definicija nije osjetljiva na mjeru C_r koja nije primjerena za mjerenje je li mreža small-world jer slučajni graf nema svojstva grupiranja. Vrijednosti koeficijenta ω ograničene su na interval između -1 i 1 bez obzira na veličinu mreže. Za vrijednost oko 0 može se smatrati da je mreža small-world što znači da je $L \approx L_r$ i $C \approx C_l$. Pozitivne vrijednosti ukazuju na to da graf ima više sličnosti sa slučajnim grafom, dok negativne na to da je graf pravilnijeg, rešetkastog oblika [6].

Posljednja mjera koja kvantificira small-world mjeru normalizira koeficijent grupiranja i duljinu puta mreže relativno u odnosu na karakteristike ekvivalentne

rešetkaste i slučajne mreže. Small World Index (SWI) računa se na sljedeći način:

$$SWI = \frac{L - L_l}{L_r - L_l} \cdot \frac{C - C_r}{C_l - Cr} \tag{2.5}$$

Mjera ima interval rezultata između 0 i 1. Što je bliže 1 to je više vjerojatno da je mreža small-world. Vrlo je vjerojatno da ne postoji mreža koja bi imala SWI = 1, ali ideja mjere je izmjeriti small-world svojstvo na način koji bi teoretski činio mrežu idealnom small-world mrežom gdje vrijedi da je $C \approx C_l$ i $L \approx L_r$.

3. Algoritmi otkrivanja društvenih zajednica

Ključan dio u pronalasku društvenih zajednica u društvenim mrežama su algoritmi koji ih otkrivaju. Oni moraju biti pouzdani i učinkoviti, ali se i izvršavati u prihvatljivom vremenskom okviru. Algoritmi se testiraju na brojnim skupovima podataka uz prikladne evaluacijske mjere kako bi se zaključilo u kojim uvjetima koji algoritam daje najbolje rješenje.

Grafove koji predstavljaju društvene zajednice teško je prikazati u ravnini ako teže stvarnim veličinama koje se kreću u tisućama čvorova, a često i mnogo više, što znači da se ne može iz ljuske perspektive odrediti kako bi dobar raspored zajednica izgledao. To znači da su algoritmi koji pronalaze društvene zajednice nenadzirani algoritmi koji sami, bez primjera za učenje i unaprijednog znanja o njima pokušavaju pronaći rješenje. U društvenim mrežama algoritmi koriste topološke karakteristike i specifičnosti koje posjeduju ovakvi tipovi mreža.

Dvije važne tehnike na kojima se temelji većina algoritama su particioniranje i grupiranje. Particioniranje grafova je proces u kojem se graf dijeli na unaprijed određeni broj manjih komponenti pomoću određenog svojstva. Svojstvo koje se može iskoristiti je minimalni rez. Ono se koristi tako da se graf podijeli na dva ili više razdvojenih podgrafova, a veličina reza koja se pokušava minimizirati je broj bridova koje je potrebno ukloniti da bi to ostvarili. Potrebno je odrediti i svojstvo koje bi odredilo veličinu komponenti kao primjerice minimalan ukupan stupanj vrhova kako bi se dobila rješenja koja imaju smisla. Zbog takvih zahtjeva ovakav pristup najčešće nije prihvatljiv jer broj zajednica nije moguće unaprijed odrediti. Grupiranje je proces u kojem se entitete koji imaju zajedničke karakteristike svrstava u iste grupe. Pronalaženje grupa može dati informacije o skrivenim značajkama, vezama i svojstvima članova te koliko su međusobno čvrsto povezani. U hijerarhijskom grupiranju stvara se hijerarhija među zajednicama. Proces se može odvijati na dva načina, aglomerativni ili divizivni. U aglomerativnom na-

činu se koristi pristup koji ide od dna prema vrhu te se određeni čvor dodaje drugim sličnim čvorovima te se koristi određeni kriterij sličnosti. U divizivnom načinu veće grupe dijele se na manje uz korištenje određene mjere koja govori koliko je dobra trenutačna podjela prema kojoj će se odrediti konačan rezultat.

3.1. Girvan-Newmanov algoritam

Veliko zanimanje i rast aktivnosti znanstvene zajednice u području društvenih mreža potaknuo je rad Girvana i Newmana iz 2002. godine.

4. Skupovi podataka

...

5. Programsko ostvarenje

...

6. Vrednovanje i rezultati

...

7. Zaključak

...

Literatura

[1] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5): 75–174, 2010.

[2] Santo Fortunato i Darko Hric. Community detection in networks: A user guide. *Physics reports*, 659:1–44, 2016.

[3] Andrea Lancichinetti i Santo Fortunato. Community detection algorithms: a comparative analysis. *Physical review E*, 80(5):056117, 2009.

[4] R Duncan Luce i Albert D Perry. A method of matrix analysis of group structure. *Psychometrika*, 14(2):95–116, 1949.

[5] Anamari Nakić i Mario Osvin Pavčević. *Uvodna poglavlja u teoriju grafova*. UNIZG-FER, 2019.

[6] Qawi K Telesford, Karen E Joyce, Satoru Hayasaka, Jonathan H Burdette, i Paul J Laurienti. The ubiquity of small-world networks. *Brain connectivity*, 1(5):367–375, 2011.

Usporedba algoritama otkrivanja zajednica u društvenim mrežama

Sažetak

Sažetak na hrvatskom jeziku.

Ključne riječi: Ključne riječi, odvojene zarezima.

Title on english

Abstract

Abstract.

Keywords: Keywords.