SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 3957

DETEKCIJA ZAJEDNICA U DRUŠTVENIM MREŽAMA

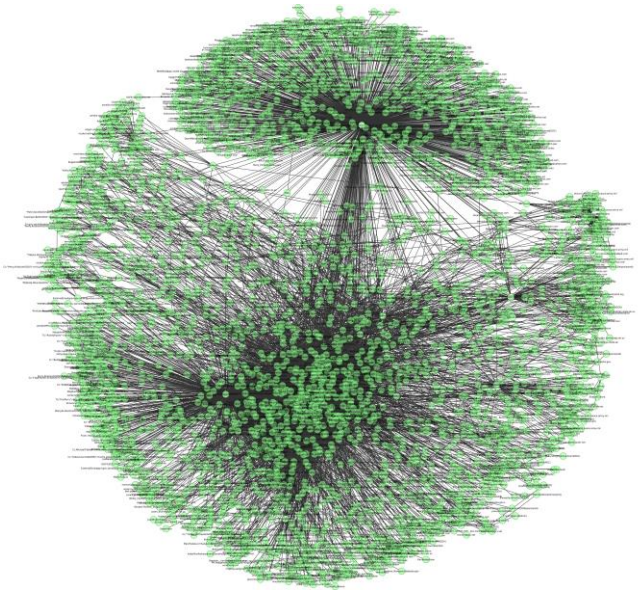
Petar Ilijašić

Zagreb, lipanj 2015.

Sadržaj

1. Uvod

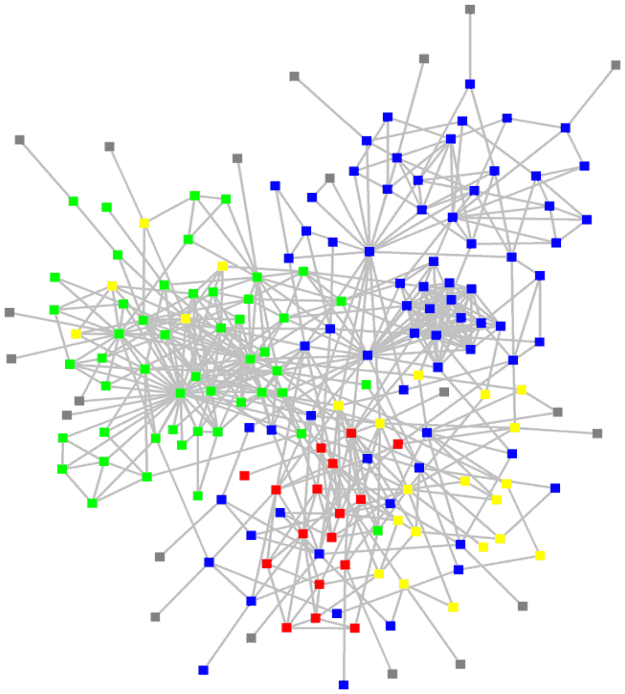
Društvene mreže su se pojavile početkom 90-ih godina 20. stoljeća. Njihova prvotna svrha se očitovala u obliku *chat-boxa,* kako bi korisnici mogli međusobno se povezati i komunicirati o temama ovisne o njihovim zajedničkim interesima. Tijekom vremena područje društvenih mreža se eksponencijalno širilo. Svrha društvenih mreža više nije bila samo komunikacija, s vremenom su se društvene mreže dotakle područja marketinga, pružanja brojnih usluga koje bi korisniku omogućile razvoj vlastitog posla. Unatoč tom razvoju, korijen pojma „*društvena mreža“* je ostao isti. Kada netko spomene društvene mreže, prva asocijacija je komunikacija s drugim korisnicima.



Slika 1.1 Grafički prikaz složene društvene mreže

Danas su društvene mreže postale jedan od glavnih faktora komunikacije u svijetu, bez kojeg komunikacija ne bi bila zamisliva današnjim generacijama.Među društvenim mrežama, najpoznatije su *Facebook, Twitter* i *Instagram.* Na slici 1.1 se nalazi primjer kompleksne društvene mreže, koja je preuzeta sa *Twitter* računa. Nakon što su te društvene mreže 2004. preuzele vodstvo, društvene mreže su se

popularizirale. Kako društvene mreže povećavaju svoj opseg, tako se količina podataka povećava. Svakodnevno putem raznih poslužitelja prolazi veliki broj podataka od korisnika tijekom međusobne komunikacije. Ukoliko se promijeni perspektiva, tako da se ne gleda društvena mreža kao korisnik, nego iz perspektive matematičara, može se doći do brojnih zanimljivih zaključaka. Veliki broj informacija se može dobiti iz većeg opsega podataka na društvenim mrežama.



Slika 1.2 Grafički prikaz „jednostavnije“ društvene mreže

Na temelju tih podataka mogu se prepoznati brojne zajednice korisnika koji su povezani međusobno bilo kojom relacijom. Najpoznatija relacija je „prijateljstvo“ na društvenoj mreži *Facebook*. Međutim detaljnijom analizom podataka mogu se uočiti brojne druge vrste zajednica. Rad je podijeljen na dva dijela. Prvi dio se sastoji od teorije koja je potrebna da bi se lakše shvatila primjena u znanstvenom radu. U prvim poglavljima je objašnjena definicija društvene mreže te pojmovi kao što su klasteriranje, komponente grafa društvene mreže i slično, dok će se u posljednja dva poglavlja opisati statistički alat R te njegova primjena u stvarnosti.

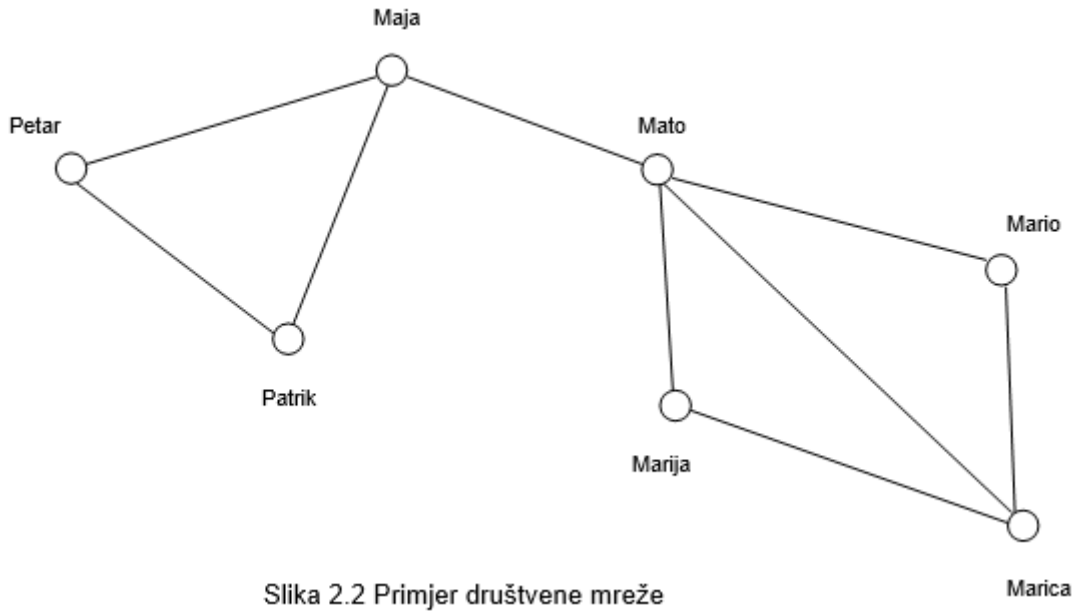
2. Definicija društvene mreže

Društvena mreža je struktura koja se sastoji od korisnika i njihovih međusobnih veza. Kada želimo definirati društvenu mrežu, kao ključne karakteristike mogu se navesti korisnici, njihove veze te pretpostavka lokalnosti zajednice. Kako se detaljnije ulazi u analizu društvenih mreža, matematički se može zaključiti da se cijela društvena mreža može prikazati kao jedan veliki graf. Graf se sastoji od točaka i rubova. Kao što se vidi na slici 2.1, svaka točka predstavlja jednog korisnika, a rub predstavlja poveznicu između te dvije točke.

Slika 2.1 Svaki korisnik društvene mreže je dio velike zajednice

Teorija grafova općenito proučava kako se ponašaju grafovi, o čemu ovisi koje su točke povezane te koja je težina veze između njih. Težina veze u području društvenih mreža predstavlja stupanj povezanosti. Veza između dvije točke, odnosno dva korisnika može biti prijateljstvo, obiteljska povezanost, poslovna i slično. Grafovi

mogu biti usmjereni i neusmjereni. Usmjereni grafovi su specifični po tome da su njihove veze jednosmjerne, odnosno graf ima početnu točku i krajnju točku. Induktivno se može zaključiti da su neusmjereni grafovi specifični po dvosmjernim vezama, bez početne i krajnje točke. Obično su društvene mreže definirane neusmjerenim grafovima. Zašto vrijedi ta tvrdnja?



Slika 2.2 Primjer društvene mreže

Ako proučimo sliku 2.2, može se uočiti graf G. Graf G predstavlja primjer male društvene mreže, koja se sastoji od nekoliko čvorova (Petar, Maja, Mario, Patrik, Matija, Marija i Marica) te njihovih međusobnih povezanosti.

Pretpostavimo da su Petar i Maja brat i sestra. Petar je Majin brat, što znači da je veza usmjerena od Petra prema Maji. Također Maja je Petrova sestra, što znači da je veza usmjerena od Maje prema Petru. Bilo bi nelogično da je krvna veza jednosmjerna, štoviše nemoguće je. Graf prikazuje sedam osoba, a veza označava odnos između dvije osobe. Prema toj statistici može biti 21 par koji bi bio povezan na neki način.

Sljedeći primjer uključuje tri osobe: Petra, Maju i Patrika. Ako je poznato da je Petar prijatelj s Majom i ako znamo da je Petar prijatelj s Patrikom, kolika je vjerojatnost da je Maja prijatelj s Patrikom? Od prijašnjeg izračuna je poznato kako

postoji 21 mogućnost da se poveže veza između dvije osobe. Također je poznato da u grafu postoji 9 rubova između dva čvora. S obzirom da je poznat odnos između Petra i Patrika te Petra i Maje, slijedi da nam preostaje 19 mogućih parova u grafu. Također slijedi da od 9 mogućih veza poznajemo dvije, što znači da je preostalo 7, što znači da je vjerojatnost poznavanja Patrika i Maje 7:19, odnosno 0.368.

Ovo je samo jedan od brojnih primjera koji se mogu proanalizirati na ovakvoj jednostavnoj mreži koja se sastoji od samo 7 čvorova, odnosno korisnika. U stvarnom svijetu broj čvorova broji u milijunima. Broj korisnika ovisi o vrsti mreže, trendu koji je aktualan te vrsti organizacije. Kao primjer se može navesti *Windows Live Messenger*, koji je bio među vodećim poslužiteljima za međusobnu komunikaciju. Broj korisnika je bio velik, što implicira na veliki broj čvorova i povezanosti. Međutim, dolaskom društvene mreže Facebook, broj korisnika odnosno čvorova u grafu mreže za *Windows Live Messenger* se postupno smanjivao, dok je broj korisnika na Facebooku porastao.

2.1 Vrste društvenih mreža

Kako ne bi došlo do pogreške u razumjevanju, društvene mreže nisu jedini izvor podataka na temelju kojih se može generirati graf korisnika i veza. S obzirom na vrstu mreža, može se navesti nekoliko kao što su:

- mreža elektroničke pošte;
- telefonska mreža;
- poslovna mreža;
- biološka mreža;
- informacijska mreža.

Navedene mreže predstavljaju samo dio brojnih mogućnosti koje se mogu iskoristiti za analizu podataka.

2.2 Mreža elektroničke pošte

Pravilnosti u komunikaciji korisnika se mogu uočiti u mreži elektroničke pošte. Kao što je prikazano na slici 2.3, mreža elektroničke pošte je jedan od osnovnih komponenti društvenih mreža koji podupire sve ostale vrste društvenih mreža. Primjerice kao čvorovi mreže se mogu deklarirati adrese elektronske pošte svakog pojedinca, dok se kao veza može definirati svaki kontakt između određene dvije adrese elektronske mreže. Kao i kod svakog grafa, ovdje je moguće odrediti usmjerenost.

Slika 2.3 Elektronička pošta je jedan od temelja društvene mreže

Da bi se detaljnije provela analiza grafa, moguće je izolirati faktor nepotrebne pošte (engl. *spam*). Pod takvom vrstom izoliranja u grafu vrijedi pretpostavka da se bilo koja veza koja je jednosmjerna, odnosno da je s jedne adrese poslana poruka na drugu, bez povratnog odgovora, automatski definira kao nepotrebna pošta. Također je moguće definirati težinu veze između dva entiteta. Težina veze

u ovom slučaju se može definirati kao učestalost komunikacije između dva korisnika. Ovaj faktor je od velike važnosti pri detekciji zajednica u društvenim mrežama. Logično je da osobe koje često funkcioniraju tvore neku vrstu manje zajednice, bila ona poslovna, prijateljska, obiteljska i slično.

2.3 Telefonska mreža

Princip kod telefonskih mreža je sličan mreži elektroničke pošte. Svaka točka u grafu predstavlja jedan telefonski broj, a veza označava komunikaciju između dva telefonska broja. Ukoliko korisnik razgovara s telefonskim automatima, pretpostavka je da je veza jednosmjerna. Međutim standardna interakcija između dvije osobe je dvosmjerna.

Slika 2.4 Telefonska mreža je od samog početka postaja jedna od vodećih načina komunikacije

Težina svake veze u grafu se može mjeriti prema učestalosti komunikacije između dva korisnika. Prema težini veze moguće je odrediti vrstu zajednice kao što su obitelj, poslovna zajednica ili članovi nekog kluba.

2.4 Poslovna mreža

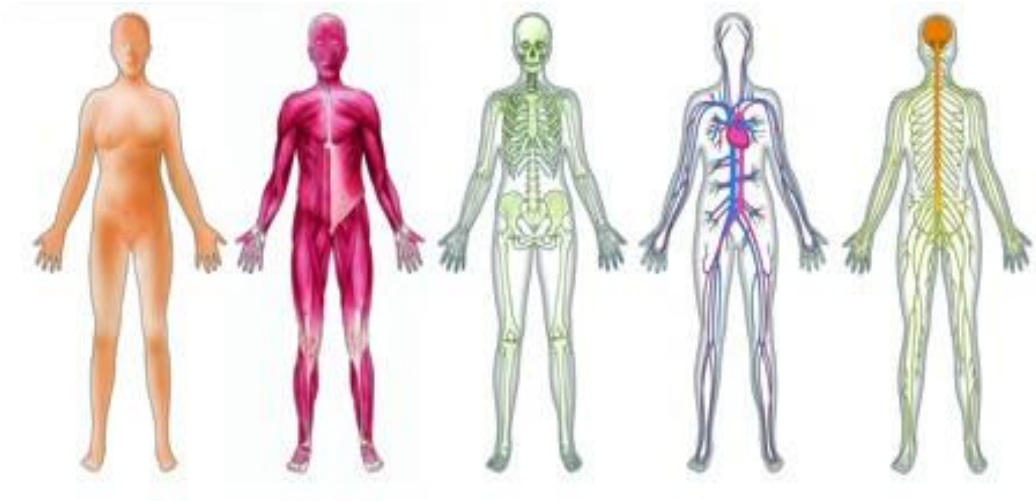
Poslovna mreža se može definirati iz različitih perspektiva. Kao primjer se može navesti poslovne kolege koji rade na projektu iz područja razvoja programske potpore. Svaki pojedini programer predstavlja jednu točku. Programeri mogu biti povezani na temelju izvornih kodova koje kreiraju u zajedničkom području rada, dok skup svih veza i točaka označava cjelokupni projekt na kojem programeri rade.

Slika 2.5 Poslovni projekti su nemogući bez kontinuirane komunikacije

Općenito bilo koja korporacija predstavlja jednu veliku zajednicu koja se sastoji od klastera. Kao što je navedeno u primjeru jedan klaster predstavlja programere koji rade na projektu, tako se može definirati zajednica na većem tržištu. Kao primjer se može navesti korporacije čije je područje rada nafta, korporacije čije je područje rada programsko inženjerstvo i slično.

2.5 Biološka mreža

Mreža ne mora biti društvena niti zahtjevati korisnike da bi se mogle detektirati zajednice u njoj. Ljudsko tijelo je samo po sebi jedna vrsta mreže koja se sastoji od zasebnih klastera. U ovom slučaju svaki klaster je jedna zajednica tkiva koji imaju određenu funkciju. Kao primjer možemo navesti živčani sustav, probavni sustav, krvožilni sustav i slično.



Slika 2.6 Ljudsko tijelo čini mrežu sustava koji međudjeluju

Ljudsko tijelo je dokaz da mreža ne može funkcionirati bez pojedinih klastera, te da na temelju funkcionalnosti pojedine veze mogu biti povezane u podzajednice ili klastere. Svaki klaster se sastoji od zasebnih dijelova koji međusobno surađuju. Primjerice živčani sustav se sastoji od neurona. Neuroni su međusobno povezani, komuniciraju i čine određenu ulogu. Da nema tog klastera, cjelokupna struktura ne bi bila funkcionalna. Kada bi se to prikazalo grafički, svaki neuron bi bio jedan čvor, dok bi veza između neurona bila jedan rub u grafu. Detaljnijim proučavanjem uloge pojedinog neurona, klaster bi se mogao podijeliti na dodatne klastere.

2.6 Informacijska mreža

Svaka informacija predstavlja jednu komponentu u cjelokupnoj mreži informacija koje se pružaju korisniku. Informacije mogu biti povezane s obzirom na njihov sadržaj, organizaciju koja ih pruža i slično.

Slika 2.7 Informacijska mreža čini osnovu protoka podataka

Svaka informacija predstavlja komponentu, a skup takvih komponenti matematički se prikazuje kao veliki graf. Internet se sastoji od velikog broja podataka. Svi podaci imaju određenu pripadnost nekom klasteru. Primjerice jedan klaster može biti skup svih telefonskih brojeva na svijetu. Kada bi se kriteriji promijenili, taj klaster može biti podijeljen na dodatne klastere kao što su primjer klasteri telefonskih brojeva, sortirane po državama. Drugi primjer klasteriranja informacija su adrese. Za svaku adresu se može odrediti pripadnost pojedinom klasteru na temelju toga u kojem se gradu nalazi ili selu. Mogućnosti klasteriranja informacija su beskonačne.

3. Komponente društvene mreže

Da bi se olakšalo shvaćanje metoda koje omogućuju detekciju zajednica u društvenim mrežama, potrebno je upoznati ključan pojam kao što je klaster (eng. *cluster*). Svaka kompleksna mreža sadrži točke koje su međusobno povezane te se na temelju kojih može odrediti razlog zašto su povezane baš na taj način. Pitanje je radi li se o obitelji, prijateljstvu, poslovnom projektu i slično. Svaki skup točaka koji je povezan na određen način da čini grupu se naziva klaster.

3.1 Konzistentni dijelovi društvene mreže

Na temelju klasteriranja grafa je moguće detektirati zajednice u društvenim mrežama. Osim klastera postoje:

- dijade;
- trijade;
- klike (eng. *Cliques*);
- klanovi;

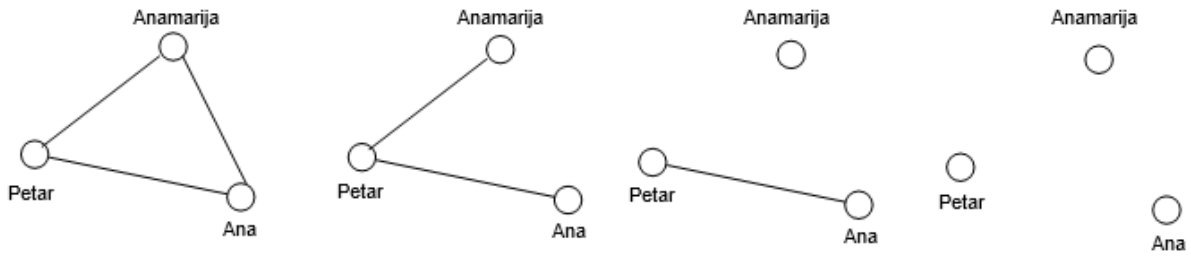
3.2 Dijade

Dijada je društvena grupa koja se sastoji od dva člana. Ova grupa predstavlja najslabiju komponentu u društvenoj mreži, jer zahtjeva obostranu interakciju. Nestankom interakcije od strane bilo koje od dva člana veza prestaje.

Slika 3.1 Dijada

3.2 Trijada

Za razliku od dijade, trijada ima dodatnog člana. Trijade mogu biti povezane na različite načine. Trijada može biti otvorena i zatvorena, također je moguće da su spojena dva čvora, a treći nije ili da čvorovi nisu povezani uopće.

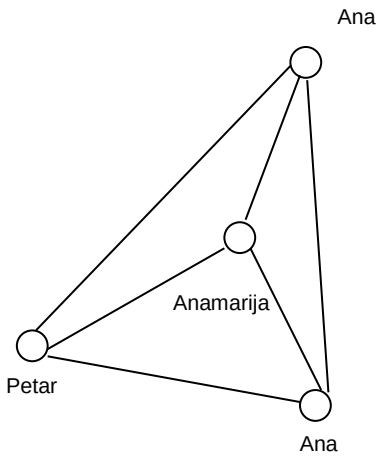


Slika 3.2 Trijada (zatvorena, otvorena, povezani par, nepovezana)

Na slici 3.2 je prikazan primjer trijade, a navedeni su članovi tima Petar, Ana i Anamarija. Ukoliko su u konstantnoj komunikaciji, vrijedi potpuna povezanost, ukoliko Anamarija prestane biti u kontaktu sa Anom, njihova veza se eliminira. Ukoliko je Anamarija odlučila napustiti tim, sve veze koje potječu od nje u toj grupi prestaju postojati. Ukoliko je tim završio projekt i više ne žele raditi zajedno, sve veze iz tog projekta prestaju postojati.

3.3 Klika

U smislu društvene mreže, klika predstavlja grupu ljudi koja je međusobno usko povezana, ali nije povezana sa vanjskim svijetom. Matematički, klika se definira kao maksimalno završen podgraf u kojem su sve točke u direktnoj vezi sa ostalima. Atribut maksimalno označava karakteristiku da ne postoje drugi čvorovi koji mogu biti dodani ovoj kliki bez da se smanji razina njezine povezanosti.



Slika 3.3 Prikaz jednostavne klike od 4 osobe

Klika se sastoji od više zatvorenih trijada koje se međusobno preklapaju te nasljeđuju mnoge karakteristike zatvorenih trijada. Kako bi klika opstala, svaki član klike mora biti biti skladna i donositi konzensus, kao što je prikazano na slici 3.3. Radi ove karakteristike lako se može primjetiti kako različite klike nisu jednake u svojim namjerama.

4. Metode pronalaženja zajednica

4.1 Pojam *Betweeness*

Pojam *betweeness* je važan pri analizi grafa društvene mreže, a označava sljedeće:

- broj čvorova između kojih se čvor nalazi;
- koliko često se čvor pojavljuje na najkraćem putu u mreži;
- jednaka je broju najkraćih puteva koji prolaze kroz čvor podijeljenom sa svim najkraćim putevima na mreži;
- normalizira se tako da je najveća vrijednost 1;
- pokazatelj gdje bi se mreža raspala, odnosno koji čvorovi bi bili otkinuti ako nestane dio čvora.

Za analizu svake osobe koja predstavlja točku na grafu, važan je faktor udaljenosti. Udaljenost je računa ukoliko su veze između dva entiteta označene nekom određenom težinom. Ta informacija je važna jer se na temelju udaljenosti provode algoritmi optimizacije i pronalaska najboljeg puta od točke A do točke B.

Grafove velikih društvenih mreža je moguće klasterirati, a klasteriranje se provodi pomoću nekoliko metoda koje su se pokazale vrlo korisnima. Općenito klasteriranje se može podijeliti na dva osnovna načina: hijerarhijsko i pomoću pridruživanja točaka.

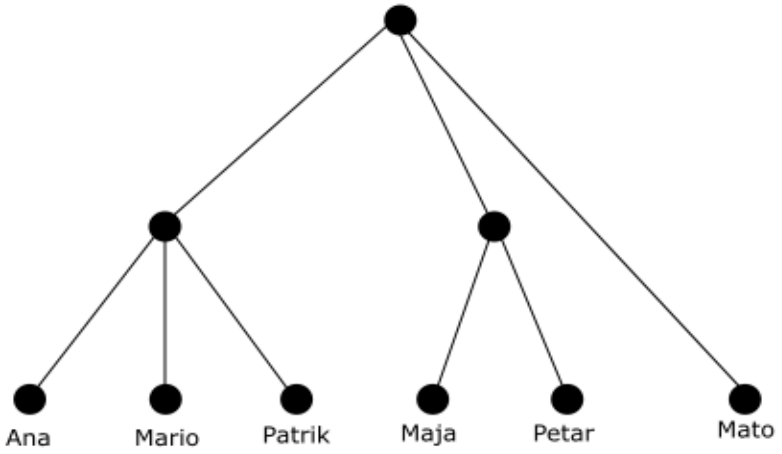
4.2 Hijerarhijsko klasteriranje

Na početku hijerarhijskog klasteriranja se uzima pretpostavka da je svaka točka u grafu zapravo u zasebnom klasteru. Nakon toga rekombinacijom točaka nastaju veći klasteri. Algoritam je sljedeći:

```
dok(nije vrijeme za prestanak){
     odaberi najbolja dva klastera za spajanje;
     kombiniraj ta dva klastera u jedan veći klaster;
} kraj;
```

Ispis 4.1 Algoritam za hijerarhijsko klasteriranja

Algoritam samo po sebi daje naziv hijerarhijski, jer uzima manje klastere i spaja u veće, pri čemu se odabirom dva klastera biraju oni koji imaju najmanju međusobnu udaljenost ***d(entitet1, entitet2)***. Primjer na slici 4.1 objašnjava kako funkcionira algoritam.



Slika 4.1 Hijerarhijsko klasteriranje

Donja razina predstavlja zasebne entitete. Svakom razinom se nanovo definira bliskost, odnosno udaljenost između entiteta prilikom čega se entiteti međusobno spremaju u veće klastere. Ana, Mario i Patrik zajedno idu na isti fakultet, Maja i Petar zajedno rade na istom poslu, a Mato je u rodnoj vezi sa svim entitetima u grafu.

Na drugoj razini hijerarhije, Petar i Maja su se spojili u jedan veći klaster. Na drugoj razini se nalaze tri klastera. U jednom klasteru su Ana, Mario i Patrik, koji će se nadalje zvati klaster A. U drugom klasteru su Maja i Petar, odnosno to je klaster B. A treći klaster C je Mato. Sva tri klastera se žele spajati dalje radi optimalnosti. Da bi

se to napravilo, potrebno je izračunati udaljenosti između klastera A, B i C te definirati udaljenosti između njih. Udaljenost između grupe A i B je obično najveća udaljenost između elemenata svake grupe, koje se zove potpuno vezno grupiranje:

$$\max_{x \in \mathcal{A},\, y \in \mathcal{B}} d(x, y) \qquad (1)$$

Također može biti najmanja udaljenost ili pojedinačno vezno grupiranje:

$$\min_{x \in \mathcal{A},\, y \in \mathcal{B}} d(x, y) \qquad (2)$$

Ili srednja udaljenost između elemenata svake grupe, koja se računa na temelju kardinalnosti:

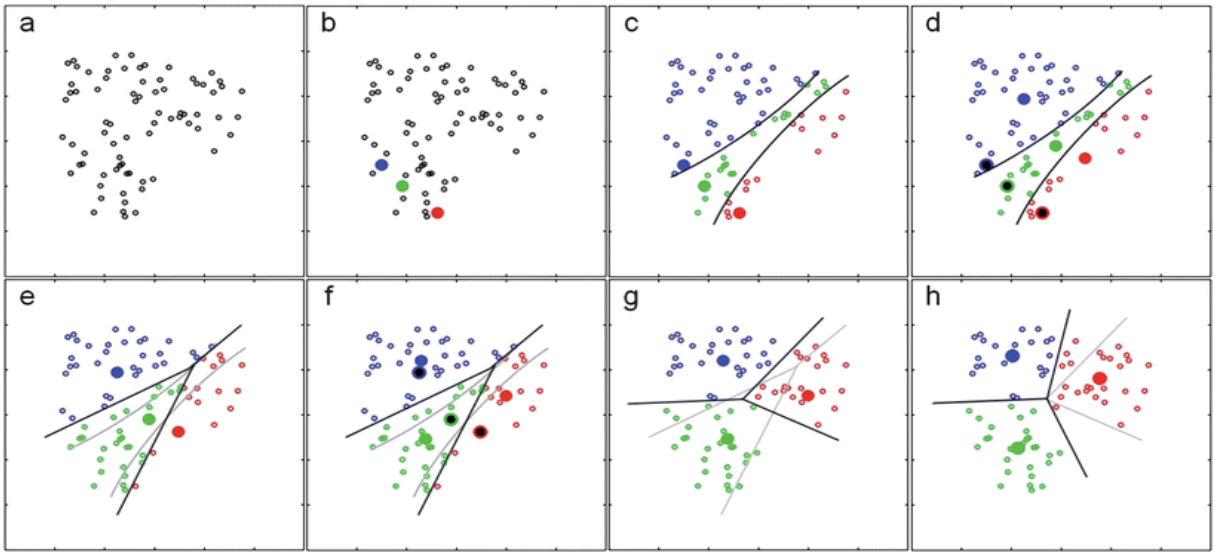
$$\frac{1}{\operatorname{card}(\mathcal{A})\operatorname{card}(\mathcal{B})} \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} d(x, y) \qquad (3)$$

4.3 Algoritam k-sredina

Algoritam se razlikuje od hijerarhijskog po tome što umjesto pretpostavke da je svaka točka sama po sebi klaster, prvo se nasumično definira k točaka tako da svaka bude u zasebnom klasteru. Svaka ta točka predstavlja centroid klastera. Algoritam glasi:

```
Za svaku preostalu točku p čini sljedeće{
     pronađi centroid prema kojem je točka p najbliža;
     dodaj točku p u klaster tog centroida;
     rekalkuliraj položaj centroida tog klastera uz p;
     ponovi radnju dok podaci ne konvergiraju;
} kraj;
```

Ispis 4.2 Algoritam k-sredina



Slika 4.2 Proces algoritma k-sredina

Na slici 4.2 je prikazan postupak klasteriranja korištenjem algoritma k-sredina. Na slici a) je prikazan skup čvorova koji se nalaze na nasumičnim položajima. Na slici b) su prikazani tri nasumično deklarirana centroida. Nakon početne deklaracije centroida ostali čvorovi se opredjeljuju pojedinom centroidu, ovisno o tome kojem je najbliži, što je prikazano na slici c). Na slici d) je prikazana rekalkulacija položaja centroida s obzirom na čvorove koji su u njegovom klasteru. Rekalkulacija se mora provoditi jer mora vrijediti pravilo da čvorovi moraju biti ekvidistantni s obzirom na centrosom. Postupak se provodi do potpune faze konvergencije, koja je prikazana na slici h). Postupak sam po sebi ne garantira optimalno klasteriranje. Razlog tome je početna nasumičnost položaja centroida, što znači da bi ponovnom provedbom ovog algoritma rezultat možda bio drugačiji od prethodnog.

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 \qquad (4)$$

Klasteriranje je optimizirano ukoliko je postignuta minimalna suma kvadrata udaljenosti između centroida i čvorova unutar svakog klastera.

Kao alternativa K-means algoritmu pruža se algoritam QT klasteriranje. QT klasteriranje ne zahtjeva unaprijedno deklariranje broja grupa. Prvi korak ovakvog načina klasteriranja je definiranje kandidatne grupe za svaki pojedinačni entitet u grafu uključivanjem najbližeg entiteta sve dok se ne dosegne prag udaljenosti. Pomoću potpune povezanosti se računa udaljenost između entiteta i grupe entiteta. Kandidatna grupa sa najviše entiteta se sprema kao prva prava grupa, dok se svi entiteti eliminiraju iz daljnje analize. Algoritam se provodi na isti način uz uzastopno smanjivanje broja entiteta.

4.4 Raspršeno grupiranje C-means

Algoritam K-means općenito prikazuje grupiranje podatke uz relativno zadovoljavajuću aproksimaciju. Specifičnost kod algoritma je ta da je svaka točka pridodijeljena isključivo jednoj grupi, što je problematično jer točaka na rubu grupe ili blizu druge grupe ne mora biti jednako mnogo u grupi kao točaka u središtu grupe. U raspršenom grupiranju svakoj točci nije strogo deklarirana pripadnost određenoj grupi, nego svaka točka ima stupanj pripadnosti određenoj grupi. Za svaku točku x postoji koeficijent koji daje stupanj pripadnosti k-toj grupi $\mathbf{u_k(x)}$. Za zbroj koeficijenata mora vrijediti sljedeći relacija:

$$\forall x \quad \sum_{k=1}^{\text{num. clusters}} u_k(x) = 1. \tag{5}$$

Slično kao i kod algoritma k-sredina, potrebno je izračunati centroid grupe koji se računa kao sredina svih točaka s težinom na njihovom stupnju pripadnosti grupi, odnosno

$$\text{center}_k = \frac{\sum_x u_k(x)x}{\sum_x u_k(x)}. \tag{6}$$

Stupanj pripadnosti određenoj grupi je obrnut udaljenosti od grupe

$$u_k(x) = \frac{1}{d(\text{center}_k, x)}, \tag{7}$$

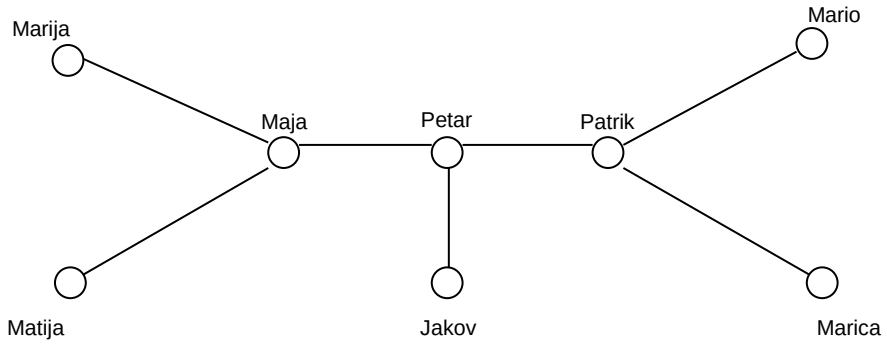
Onda se koeficijenti normaliziraju i zapliću s realnim parametrom $m > 1$ tako da je njihov zbroj 1. Dakle,

$$u_k(x) = \frac{1}{\sum_j \left(\frac{d(\text{center}_k, x)}{d(\text{center}_j, x)} \right)^{\frac{1}{m-1}}} \tag{8}$$

Koraci za C-means algoritam su sljedeći:

1. Odabere se broj grupa;
2. Dodijeli se nasumice svakoj točki koeficijente za pripadnost grupama;
3. Ponavlja se sve dok algoritam nije konvergirao (to jest, koeficijentska; promjena između dva ponavljanja nije veća od ϵ, dani prag osjetljivosti):
    a. Izračuna se centroid za svaku grupu upotrebom gornje formule;
    b. Za svaku se točku izračunaju njeni koeficijenti pripadnosti grupama upotrebom gornje formule;

Algoritam raspršenih *c*-sredina jednako smanjuje međugrupnu različitost, ali ima iste probleme kao *k*-sredine, minimum je lokalni minimum, a rezultati ovise o početnom odabiru težina.



Slika 4.2 Primjer društvene mreže za određivanje pojavljivanja čvora

Slika 4.2 prikazuje jednu skupinu korisnika društvene mreže. Korisnici nisu potpuno povezani međusobno, već preko drugih osoba. Ukoliko Marija želi predati informaciju Marici, Marija će morati proslijediti informaciju preko Maje, Petra i Patrika. Ako Jakov želi poručiti Mariu neku važnu informaciju, on će morati preko Petra i Patrika prenjeti poruku. Može se primjetiti da je Petar ključna osoba koja prenosi većinu informacija među korisnicima. Stoga se matematički može reči da je on čvor kroz koji se najčešće prolazi na najkraćem putu. Ukoliko se simuliraju svi prolazi, dolazi se do zaključka kako će Petar prenjeti 15 puta informaciju, a Petar i Maja 11 puta. Ostali korisnici se nalaze na početku grafa, stoga je logično da oni mogu samo inicirati poruku, ne ju prenositi.

5. Model SimRank

Model SimRank općenito predstavlja mjeru za sličnost između dva objekta koji su međusobno povezani. Sličnost dva objekta se temelji na njihovim vezama s ostalim objektima, odnosno se može reći kako se dva objekta mogu smatrati sličnima ukoliko su referencirani od strane sličnih objekata. Kao što je objašnjeno u prijašnjim poglavljima, društvene mreže mogu biti različite. Kao jedan od primjera je navedena informacijska mreža, gdje je klasteriranje dokumenata jedan od ključnih karakteristika za uvođenje pravilne strukture mreže. World Wide Web je jedan od osnovnih primjera potrebe aplikacije za korištenjem ovog modela. U ovom slučaju potrebno je pronalaziti dokumente slične po sadržaju u slučaju da korisnik želi istražiti neko određeno područje rada. Dva su dokumenta u srodstvu ukoliko su povezani hipervezom. Sličan pristup može biti prilikom određivanja zajedničkih svojstava znanstvenim dokumentima na temelju područja znanosti koji se obrađuju.Također se algoritam osim u dokumentima može poslužiti za bilo koju potrebu gdje korisnik želi dobiti što više međusobno povezanih podataka. Ukoliko korisnika zanima neka određena odjeća, pomoću modela SimRank bi vrlo lako dobio sve potrebne podatke o toj odjeći te odjeći sličnoj traženoj. Model SimRank može poslužiti u klasteriranju, odnosno pronalaženju društvenih zajednica te se obično prikazuje pomoću grafova. Ključno je znati kako se model SimRank može primjeniti isključivo na strukturiranim sadržajima, odnosno na područja u kojima postoji dovoljan broj objekata s validnim vezama između njih. Pri analiziranju društvene mreže taj uvjet je zadovoljen jer je društvena mreža općenito definirana pomoću strukturiranog grafa koji se sastoji od točaka koje su međusobno povezane.

5.1 Matematička notacija SimRank modela u teoriji grafova

Ako se pretpostavi da postoji usmjereni graf koji se sastoji od nekoliko točaka. Svaka točka predstavlja osobu, a veza između njih može biti jednosmjerna ili dvosmjerna. Neka jedna točka iz grafa bude nazvana v. Ulazni susjedi $I(v)$ točke v

su oni koji imaju vezu koja je usmjerena prema točci v, a izlazni susjedi $O(v)$ točke v su oni čija je veza usmjerena od točke v. Sličnost između objekata *a* i *b* se deklarira kao:

$$s(a,b) \in [0,1] \tag{9}$$

Ukoliko je a identičan b, to bi značilo da je sličnost između a i b jednaka 1. U ostalim slučajevima vrijedi jednakost:

$$s(a,b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s(I_i(a), I_j(b)) \tag{10}$$

gdje je **C** konstanta iz intervala <0, 1>.

Postoji mogućnost da **a** ili **b** nemaju nikakve ulazne susjede. U tom slučaju nije moguće odrediti bilo kakvu sličnost između objekata **a** i **b**, što bi značilo da je sličnost između **a** i **b** jednaka nuli. Općenito ukoliko je broj ulaznih susjeda barem jedne od dvije točke jednaka nuli, sličnost je također jednaka nuli.

Model SimRank se sastoji od jednadžbi za grad ***G,*** pri čemu se one rješavaju iteracijom kroz fiksnu točku. Neka je n broj točaka u grafu ***G.*** Za svaku iteraciju *k* se može zadržati n^2 ulaza sličnosti $s_k(*,*)$. Rezultat iteracije između točki *a* i *b* se izražava kao $s_k(a,b)$. Na temelju ulaza sličnosti lako se može izračunati $s_{k+1}(*,*)$. Račun započinje određivanjem $s_0(*,*)$, gdje je svaki $s_0(a,b)$ definiran kao donja granica trenutnog rezultata izračuna $s(a,b)$. Izračun sličnosti povećavanjem faktora *k* ne opada, odnosno za svaki k vrijedi sljedeća jednakost:

$$a, b \in V, \quad \lim_{k\to\infty} s_k(a,b) = s(a,b) \tag{10}$$

5.2 Optimizacija SimRank modela

Tijekom vremena izračun jednadžbi SimRank modela je prolazio kroz procese optimizacije, kako bi se konačno uspostavili tri načina optimizacije:

- odabiranjem osnovnih točki za izračun, pri čemu se a-priori izbjegava izračun s točkama bez ulaznih susjeda;
- memoriziranjem parcijanih suma prilikom izračuna sličnosti kako bi se pri novom izračunu između dvije slične točke mogao koristiti dio izračuna koji se prije provodio;
- postavljanje praga sličnosti kako bi se smanjio broj proračuna između parova točaka.

Korištenjem parcijalnih suma se dokazala optimizacija, odnosno smanjenje složenosti sa $O(Kd^2n^2)$ na $O(Kdn^2)$. $\mathbf{K}$ predstavlja broj iteracija, $\boldsymbol{d}$ predstavlja prosječan stupanj točaka u grafu, a *n* predstavlja broj točaka u grafu. Parcijalne sume preko $\mathcal{I}(a)$ se računaju sljedećim izrazom:

$$Partial^{s_k}_{\mathcal{I}(a)}(j) = \sum_{i \in \mathcal{I}(a)} s_k(i,j), \qquad (\forall j \in \mathcal{I}(b)) \tag{11}$$

Ukoliko računamo sličnost točke a sa određenom točkom, može se ponovno koristiti parcijalna suma u kojoj je točka $\mathbf{a}$ bila faktor.

6. Model brojanja trokuta u društvenim mrežama

Model brojanja trokuta je postao jako koristan u velikim grafovima kao što su grafovi društvene mreže. Kako se model brojanja trokuta odnosi na teoriju grafova? Pretpostavka je da imamo $\mathbf{n}$ točaka te se nasumično dodaje $\mathbf{m}$ grafova. Može se zaključiti kako će se u grafu pojaviti točke povezane na način da će činiti trokut. U cijelom grafu se nalazi $\mathbf{n}$ povrh 3 ili aproksimativno $(\mathbf{n}^3)/6$ skupova od tri točke koje bi mogle činiti trokut. Vjerojatnost povezanosti između bilo koje dvije točke u skupu je m/(n povrh 2) ili aproksimativno $2\mathbf{m}/(\mathbf{n}^2)$. Vjerojatnost da su tri točke povezane međusobno je jednaka $(8\mathbf{m}^3)/\mathbf{n}^6$. Ako se uzmu u obzir ove aproksimacije, očekivani broj trokuta u grafu iznosi $(8\mathbf{m}^3/\mathbf{n}^6)(\mathbf{n}^3/6) = 4/3(\mathbf{m}/\mathbf{n})^3$. Aproksimacija je provedena na općenitom matematičkom slučaju. Ukoliko se uzme u obzir da će se brojati trokutovi u grafu kojim je deklarirana društvena mreža, očekivani broj trokuta će biti veći jer svaka točka grafa simbolizira jednu osobu. Zašto bi broj trokuta bio veći? Ako se uzme u obzir da je osoba $\mathbf{A}$ prijatelj od osobe $\mathbf{B}$, a osoba $\mathbf{B}$ je prijatelj sa osobom $\mathbf{C}$, velika je vjerojatnost da će se osoba $\mathbf{A}$ i osoba $\mathbf{C}$ upoznati, nakon čega će se stvoriti trokut prijateljstva.

6.1 Algoritam za brojanje trokuta

Pretpostavka je da postoji graf sa n točaka i m rubova, pod uvjetom da je broj rubova veći od broj točaka. Svaka točka predstavlja broj u intervalu $[1, \mathbf{n}]$ te ima svoj stupanj povezanosti. Točka se zove *heavy hitter* ukoliko je njezin stupanj povezanosti barem jednak $\sqrt{\mathbf{m}}$. *Heavy hitter* trokut je trokut u kojem je svaka točka tog trokuta *heavy hitter*. Broj *heavy hitter* točaka nikad nije veći od $2\sqrt{\mathbf{m}}$, jer bi u suprotnom zbroj stupnjeva *heavy hitter* točaka bio veći od $2\mathbf{m}$. Kako svaki rub doprinosi stupnju dviju točaka, što bi značilo da bi moralo biti više od m rubova. Prvo je potrebno u nekoliko koraka proanalizirati zadani graf:

1. Potrebno je izračunati stupanj svake točke, pri čemu je složenost postupka jednaka $\mathbf{O(m)}$;
2. Za svaki rub je potrebno definirati indeks, dok će par točaka koje on povezuje biti njegov ključ. Na temelju toga se kreira tablica koja označava koje su dvije točke povezane kojim rubom. Složenost postupka je također jednaka $\mathbf{O(m)}$;
3. Kreira se još jedan indeks ruba sa ključem koji je jednak pojedinoj točci. Na taj način se može znati za svaku točku zasebno s kojim je sve točkama povezana.

Nakon što je provedena analiza grafa, potrebno je sortirati podatke. Prvo je potrebno poredati toče prema stupnju povezanosti. Ukoliko dvije točke imaju isti stupanj povezanosti, točke će se poredati prema njihovoj numeričkoj vrijednosti. Kao što je navedeno na početku poglavlja, postoje dvije vrste trokuta: *heavy hitter* trokuti i regularni trokuti.

6.2 Trokutovi Heavy hitter

U grafu se nalazi $\mathbf{O((\sqrt{m})}$ *heavy hitter* točaka te je moguće $\mathbf{O(m^3/2)}$ *heavy hitter* trokutova. Pomoću indeksa koji su definirani na rubovima, moguće je provjeriti postoje li sva tri ruba u $\mathbf{O(1)}$ vremenu. Prema tome bi značilo da je potrebno $\mathbf{O(m^3/2)}$ vremena kako bi se pronašli svi *heavy hitter* trokutovi.

6.3 Regularni trokutovi

Osim *heavy hitter* trokutova, u grafu postoje i regularni trokutovi. Postupak pronalaženja se bitno razlikuje od pronalaženja *heavy hitter* trokutova. Prvo je potrebno analizirati sve rubove grafa. Ukoliko graf spaja dvije *heavy hitter* točke, veza se ignorira jer to spada u domenu *heavy hitter* trokutova. Analizirajući točku A pod pretpostavkom da nije u kategoriji *heavy hitter* točaka te da su $\mathbf{B_1, B_2, B_3...B_k}$ susjedne točke od točke $\mathbf{A}$. Te točke se pronalaze na temelju indeksa koji su definirani na pojedinim rubovima. Složenost algoritma je $\mathbf{O(m^3/2)}$.

7. Primjena klasteriranja u stvarnoj društvenoj mreži

Detekcija zajednica u društvenim mrežama zahtjeva analizu i proučavanje velikog broja podataka. Osim toga, potrebno je poznavati adekvatnu teoriju kako bi rad na velikom broju podataka bio jednostavniji. U ovom poglavlju će biti prikazan način detektiranja zajednica u društvenim mrežama, kako bi se čitatelju cjelokupna teorija bila manje apstraktna.

Slika 7.1 Društvena mreža Facebook je vodeća prema broju aktivnih korisnika

Analiza će se provoditi kroz podatke sa društvene mreže Facebook, koje posjeduje autor ovog rada. Zbog vlastite znatiželje autora i znanstvenih razloga, istražiti će se način kako detektiranje zajednica funkcionira na točkama u grafu koje predstavlja osobe koje autor osobno poznaje.

7.1 *Dataset* u praktičnom smislu

Dataset je naziv za bilo koji skup podataka koji može poslužiti za prikaz korisniku, analizu podataka, provođenje sigurnosnih postavki i slično. Većina podataka su *open source*. Razlog tome jer veliki broj programera svakodnevno zahtjeva informacije kako bi njihovi sustavi funkcionirali pravilno. Brojne službene stranice omogućuju korisnicima dohvaćanje originalnih skupova podataka, kao što je edukativna stranica fakulteta Stanford. Neki od primjera skupova podataka su primjerice podaci sa servera poznatog portala Youtubea ili Facebooka. Youtube je danas jedna od najvećih multimedijskih stranica koje postoje na internetu. Server je zadužen za manipuliranjem velikim količinama podataka kao što su pretplatnici, korisnički kanali i slično. Korisnički kanali se mogu klasterirati s obzirom na područje rada u kojem je korisnik. Primjerice korisnički kanali mogu biti znanstvene prirode, glazbene, zabavne i slično.

Slika 7.2 Youtube je trenutno najpopularnija multimedijska stranica

Prema novijim istraživanjima na Youtubeu, klasteriranjem kanala se uspostavilo kako se razvilo novo područje koje ubrzano raste – *gaming* kanali. Razlog tome je velika popularnost među gledateljima koja se također evidentira te profit koji se dobiva svakim pretplatnikom.

Kada se spominje društvena mreža, prva asocijacija je društvena mreža Facebook. Facebook je nastao 2004. godine te mu je svrha bila međusobna komunikacija između studenata Harvarda. Svaki student je predstavljao jednu točku na grafu, odnosno jedan skup podataka. Međutim kako su se ostali fakulteti počeli uključivati u Facebook mrežu, mreža je sama po sebi bila sve veća te su skupovi podataka bili veći, a time i graf. Danas je na Facebook mrežu prijavljeno 1.230,000,000 aktivnih korisnika.

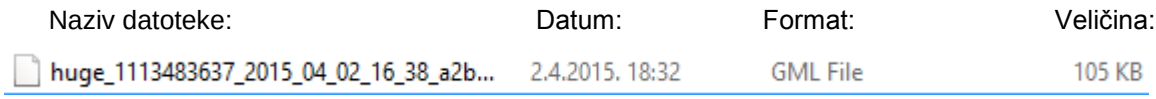
Drugi primjeri skupova podataka su oni koje drži stranica Amazon, e-Bay i slično. Na temelju svih podataka koje se svakodnevno spremaju u bazu podataka servera pojedinih portala, omogućena je moderna komunikacija između korisnika uz sve manju mogućnost pogreške.

7.2 Facebook dataset – korisnički profil

Svaki korisnički profil je jedan skup podataka na kojem se mogu provoditi detaljne analize i detekcija društvenih zajednica. Korisnik može imati veliki broj kontakata na Facebook mreži, međutim to ne znači nužno da je u istim odnosima sa svakim od njih. S nekim može biti u krvnom srodstvu, poslovnom, prijateljskom odnosu ili kao poznanik s kojim se čuo samo jednom.

Slika 7.3 Što je veći broj korisnika, potrebno je kvalitetnije klasteriranje

Na temelju više kriterija se može provesti klasteriranje nad svim osobama s kojima je korisnik u kontaktu na društvenoj mreži Facebook. Kao praktičan primjer koristi se Facebook profil autora ovog rada (privatni podaci će biti zaštićen radi sačuvanja privatnosti drugih korisnika).Kako bi se provela analiza podataka, potrebno ih je dohvatiti sa servera. Podaci su u većini slučajeva pohranjeni u GML (**G**eography **M**arkup **L**anguage).



Slika 7.4 Datoteka sa podacima o korisničkoj društvenoj mreži

Razlog zašto se koristi ovaj format je taj što je GML format normiran za prikazivanje vektorskih podataka, kako bi se olakšao prikaz geografskih lokacija i grafičkih podataka. To je u ovom slučaju korisno jer se obrađuju podaci koji predstavljaju graf društvene mreže korisnika.

```
graph
[ directed 0
    node [
        id 0
        label "Gordana Alfeldi Banova"
        sex "female"
        agerank 289
        wallcount 671
        locale "hr_HR"
    ]
    node [
        id 1
        label "Rut Pušelj"
        sex "female"
        agerank 288
        wallcount 176
        locale "hr_HR"
    ]
```

Slika 7.5 Prikaz strukture grafa društvene mreže od Facebook korisničkog profila

Na slici 7.5 je prikazan dio strukture grafa društvene mreže Facebook. Graf se sastoji od osoba s kojima je autor u kontaktu preko Facebook računa. Svaka osoba u formatu GML je definirana sljedećim zapisom:

```
node [
        id 9
        label "Petar Petrović"
        sex "male"
        agerank 280
        wallcount 535
        locale "en_GB"
]
```

Ispis 7.1 Struktura čvora koji predstavlja osobu na mreži

Pojam *node* definira jednu čvor u grafu. Jedna točka predstavlja kontakt od autora. Svaka točka je definirana jedinstvenim identifikacijskim brojem te ostalim karakteristikama kao što su spol, govorno područje, broj zapisa na „Zidu“ i slično. Sve točke su na neki način povezane. Povezanost je definirana sljedećim zapisom:

```
edge [
        source 3
        target 9
]
```

Ispis 7.2 Struktura ruba u grafu društvene mreže

Pojam *edge* predstavlja rub, odnosno vezu između između dva čvora. *Source* predstavlja izvorište veze, a *target* predstavlja destinaciju (Petar je prijatelj sa Majom). Na temelju ovih komponenti će se pomoću adekvatne programske podrške omogućiti analiza i detekcija zajednica u zadanoj društvenoj mreži.

7.3 Gephi

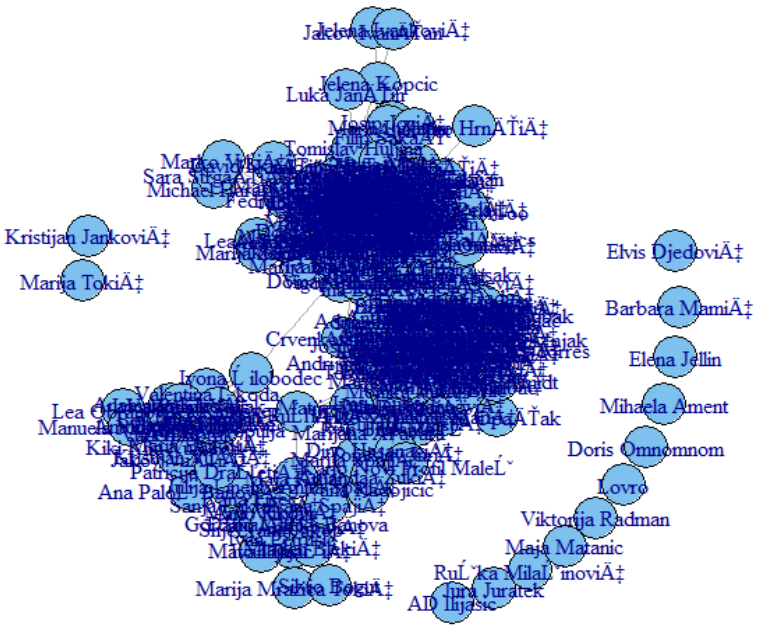
U današnje vrijeme postoji veliki broj programskih podrški za analizu društvenih mreža, kao što su *Gephi*, *UCINET*, *Pajek*, *NetMiner* i slični. Za analizu društvene mreže u ovom primjeru će se koristiti *Gephi.* Gephi je *open-source* alat za analizu društvene mreže te njenu vizualizaciju. Koristio se u Googleovim ljetnim kampovima od 2009. do 2013. godine za brojna akademska istraživanja. Svojim brojnim funkcionalnostima omogućuje detaljne proračune kao što su težina puteva u grafu, modularnost i slično.

7.4 Statistički alat R

Statistički alat R je kreiran u svrhu provođenja statističkog programiranja i grafičkog prikazivanja. Koriste ga korisnici kojima je područje rada obrada podataka i njihova analiza te statističari. Kako se povećavala količina podataka na Internetu, potreba za alatom R je također bila povećana. Jezik obavezno dolazi uz programsku potporu, koja je open source. U ovom radu će se koristiti programska podrška *RStudio*, koja je besplatna i dostupna svima koji žele učiti R. Glavni razlog korištenja *RStudia* je njegova bogata biblioteka koja sadrži veliki broj funkcionalnosti korisne programeru.

7.5 Analiza društvene mreže profila s Facebook računa

S obzirom na to da postoji veći izbor što se tiče klasteriranja zajednica, u sklopu ovog znanstvenog rada će se koristiti algoritam k-sredina. Algoritam je jednostavan za provedbu, te može čitatelju jasno prikazati na koji način funkcionira klasteriranje podataka. Na slici 7.6 je prikazan graf društvene mreže nakon što su programski implementirani dohvaćeni podaci. Svaki čvor predstavlja jednu osobu koju autor rada poznaje na društvenoj mreži Facebook te je označena imenom te osobe.



Slika 7.6. Društvena mreža korisnika prikazana grafom (sa imenima)

Radi daljnje jednostavnosti te veće preglednosti prikaza čvorova i zaštite privatnosti korisnika, imena osoba će biti zamijenjena identifikacijskim brojem čvorova. Ukoliko se analiziraju numeričke vrijednosti dohvaćenih podataka, može se primjetiti kako su oni prikazani vektorski. Razlog tome je činjenica da nam je potreban matematički oblik podataka kako bi se mogao generirati pravilni graf.

```
[289] 203 208 221 222 211 221 234 211 224 234 237 239 222 223 239 216
[305] 239 222 211 215 221 224 227 234 222 226 210 239 218 234 243 250
[321] 261 267 242 256 257 269 241 243 246 250 261 267 243 250 267 269
[337] 243 261 252 240 245 249 250 261 267 252 257 241 253 243 250 253
[353] 243 244 248 252 250 261 267 264 270 279 285 270 274 277 280 270
[369] 277 280 270 273 285 286 270 274 288 270 274 277 280 288 270 270
[385] 280  49  56  33  39  41  46  52  55  39  46  37  45  50  52  59
[401]  45  52  49  52  54  46  49  50  52  55  56  52  59  51  56  54
[417]  59  55  64  67  70  71  73  74  78  88  72  73  84  67  82  84
[433]  85  88  65  76  78  80  73  73  74  78  83  67  80  82  69  72
[449]  85  83  64  67  70  72  75  78  80  83  69  78  80  74  75  78
[465]  64  67  70  76  77  80 102 115  99 100 107 113 119 117  95 101
[481] 103 113 117 119  98  99  99 106  93  94 106  94  98  99 103  96
[497]  99 101 113  99 101 107 109 113 117 119  98  99 105  95 105 106
[513] 100 117  95 100 101 102 103 115 100 103  98  99 101 103 114 119
[529] 103 113 114  99 100 101 103 113 102 115 105 104 108 111  95  98
[545]  99 103 129 135 141 123 130 132 133 145 140 130 135 140 123 134
```

Slika 7.7 Prikaz dijela čvorova prema njihovim težinama prikazanim vektorski

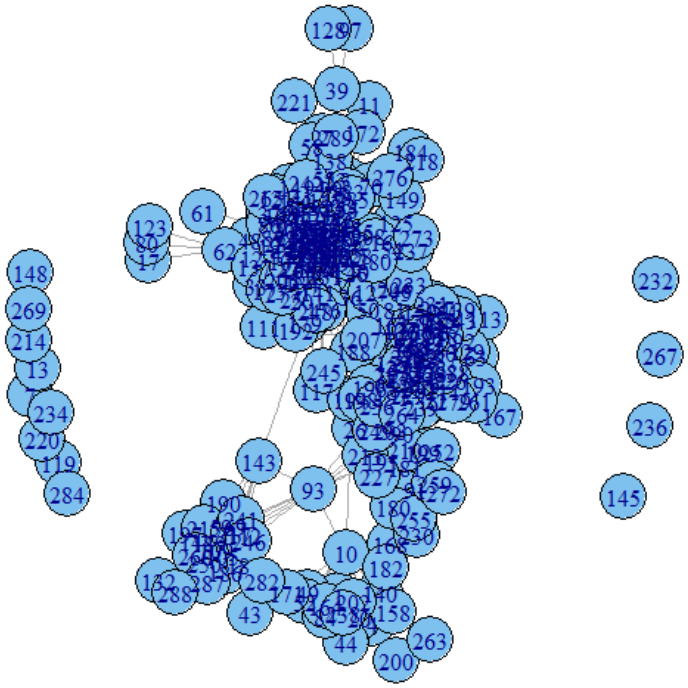
Nakon što su podaci pravilno oblikovani, klasterizacija je omogućena. Na slici 7.7 se vidi raspored točaka koje su prikazane vektorski. Svaki vektor je označen vrijednošću u uglatoj zagradi, kao na primjer [5281]. Ostale vrijednosti predstavljaju težine. Provedbom algoritma k-sredina, svaki čvor je definiran klasterom koji pripada. U ovom slučaju vrijednost k iznosi tri, odnosno u mreži grafa društvene mreže postoje tri klastera. Na slici 7.8 je prikazan dio strukturiranog zapisa čvorova prema njihovim klasterima. Svaki red predstavlja jedan vektor, koji se sastoji od čvorova u nekom klasteru.

```
[5281] 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 1 1 1 1 1 3 3 3 3 3 1 1 3 3 1 1 1 1 3 3
[5321] 3 3 3 3 1 3 1 3 3 3 1 3 3 3 3 3 1 3 3 3 1 2 1 1 3 1 1 3 3 3 2 1 1 3 3 3 3 3 3 2 3
[5361] 1 3 2 2 2 2 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 2 2 1 1 3 2 2 1 1 1 1 1 1 3 3
[5401] 3 3 3 3 1 1 1 3 3 3 3 3 2 1 1 1 3 3 3 2 2 2 2 1 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3
[5441] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 1 1 1 1 3 3 3 2 2 2 1 1 3 3 3 3 3 3 3 3 3
[5481] 3 3 3 3 3 3 1 3 3 3 3 3 1 3 3 2 1 1 3 3 3 3 2 1 3 3 3 3 3 3 3 3 3 3 3 3 2 1 1 1 1
[5521] 1 3 3 3 3 3 3 3 3 1 1 2 2 2 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2
[5561] 2 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 2 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3
[5601] 3 3 3 3 3 3 2 2 1 1 1 1 3 3 3 3 3 3 3 3 3 3 1 1 1 3 3 3 3 3 3 3 3 3 2 1 3 3 3 2 1 1
[5641] 1 3 3 3 3 3 3 3 1 1 1 1 3 3 3 3 3 3 3 1 1 1 1 3 3 3 1 1 3 3 2 2 2 2 2 1 1 1 1 1 1 1 3
[5681] 3 3 3 3 3 3 3 3 3 3 3 3 1 1 3 3 3 2 3 1 1 1 3 3 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[5721] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 1 3 3 3 3 3 3 3 3 3 2 2 1
[5761] 1 3 3 3 3 3 3 2 1 1 1 3 3 3 3 3 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3
[5801] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 1 3 2 1 3 3 3 3 3 2 1 3 3 3 3 3 3 3 1
[5841] 1 3 3 3 3 3 3 3 3 2 1 3 2 1 1 3 3 3 3 3 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3
[5881] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 3 3 3 3 2 2 2 1 1 1 1 1 1 1 1 1
```

Slika 7.8 Prikaz dijela čvorova koji su raspoređeni prema jednim od tri klastera

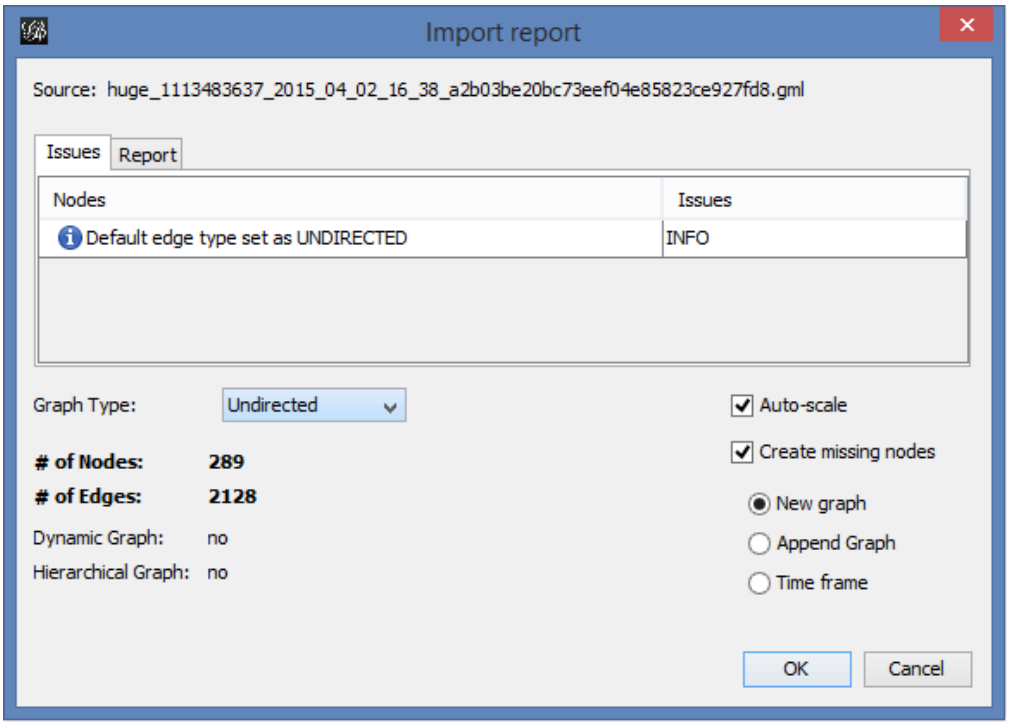
Nakon što su klasteri čvorova definirani, potrebno je provjeriti koje su vrijednosti parcijalnih suma kvadrata za svaki klaster. Prema proračunima za svaki klaster, koje ćemo nazvati **A**, **B** i **C**, sume kvadrata iznose 97040195, 93133988 i 98423486. S obzirom da k-means algoritam ne garantira optimiziranu klasterizaciju, potrebno je provjeriti točnost provedbe algoritma. Prema analizi proračuna suma kvadrata pojedinih klastera u odnosu na sumu od svih klastera, točnost klasteriranja iznosi 93%. Unatoč dobivenim rezultatima, prilikom sljedećeg proračuna rezultati ne moraju biti nužno isti, tako da će korisnik koji provede analizu dobiti vrlo vjerojatno drugačiji rezultat. Prednost statističkog alata R je njegova bogata biblioteka koja se koristila prilikom ove analize. U programskom jeziku postoji zaseban paket koji služi isključivo za analizu društvenih mreža, algoritmizaciju te vizualni prikaz podataka. Ta činjenica

znatno olakšava bilo kome koga zanima kako algoritam funkcionira da samostalno pokuša detaljno pregledati proces klasteriranja. Rezultat provedenog algoritma u ovom slučaju je prikazan na slici 7.9.



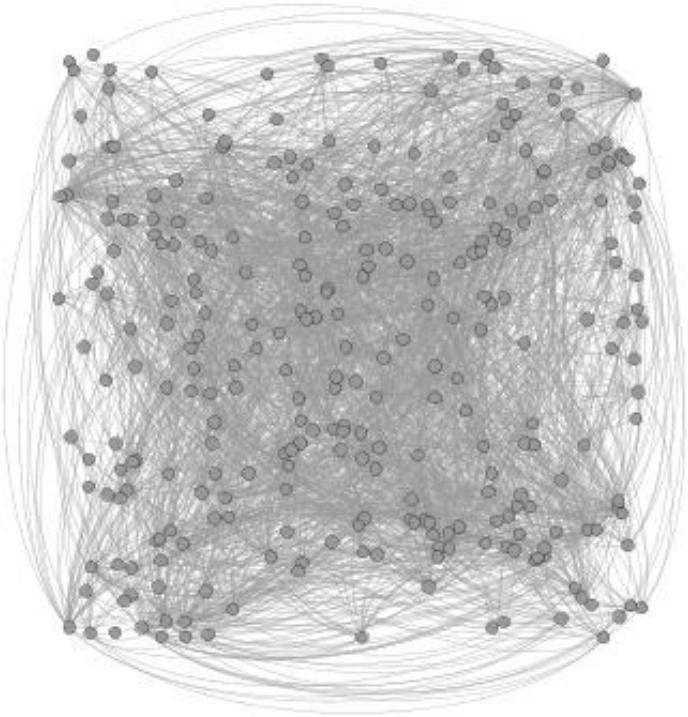
Slika 7.9 Rezultat provedbe klasteriranja prema rezultatu k-sredina

Kako bi se provela analiza podataka sa društvene mreže, nije nužno potrebno kreirati vlastite skripte u programskom jeziku R, već je moguće koristiti gotova programska okruženja. Na taj način klasteriranje mogu provoditi i korisnici koji nemaju iskustva sa tim jezikom. U sljedećem primjeru koristit će se programsko okruženje *Gephi*. U ovom primjeru se koristi inačica *Gephi 0.8.2*. Podaci koji će se analizirati su isti kao i u *RStudiu*, s tim da su čak i opširniji jer su osim numeričkih podataka svake osobe uključeni i tekstualni podaci. Samim time je jedini uvjet za uspješno provođenje klasteriranja u *Gephiu* točnost formata datoteke te korisnik ne mora nužno manualno filtrirati podatke. Jedino što je potrebno je datoteka sa prikladnim podacima koji su u formatu GML, CSV i slični.



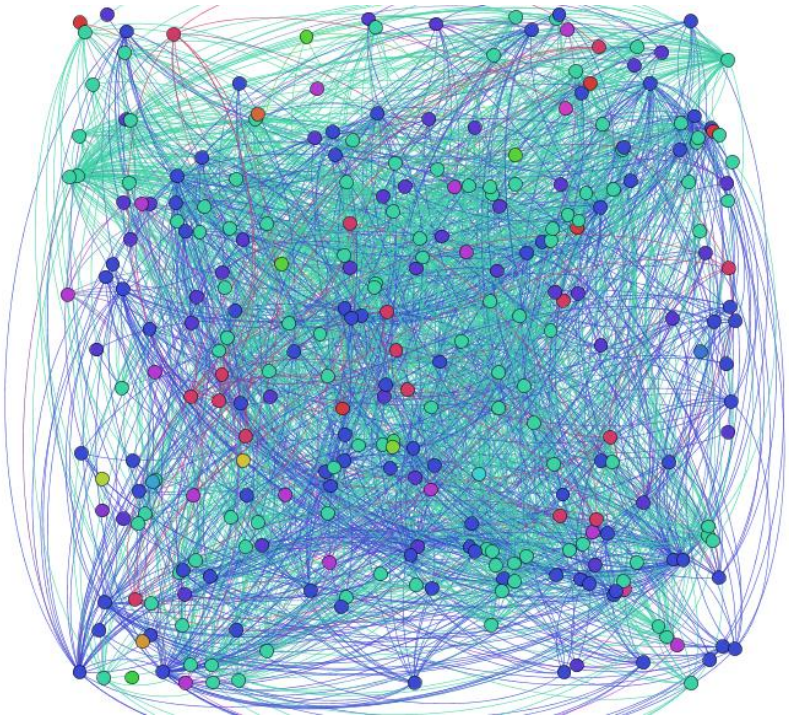
Slika 7.10 Implementacija podataka u *Gephi*

Nakon što su podaci uneseni u program, prije generiranja vizualne mreže se ispisuje specifikacija o unesenim podacima. Prema podacima se vidi kako će u mreži biti 289 čvorova, što znači da korisnik ima 289 kontakata na svom Facebook računu (moguće je da neke osobe nisu uključene u podacima zbog neaktivnosti njihovog računa ili postavkama sigurnosti). Također je poznato da su čvorovi povezani sa 2128 veza. Graf sam po sebi nije potpuno povezan, znači da veze nisu jednoliko raspoređene prema čvorovima. Dodatni kriterij je vrsta grafa s obzirom na usmjerenost, što je u ovom slučaju neusmjereni graf. Rezultat generiranja mreže je prikazan na slici 7.11, koja će biti nadalje analizirana do faze kada će klasteri biti jasno vidljivi korisniku.



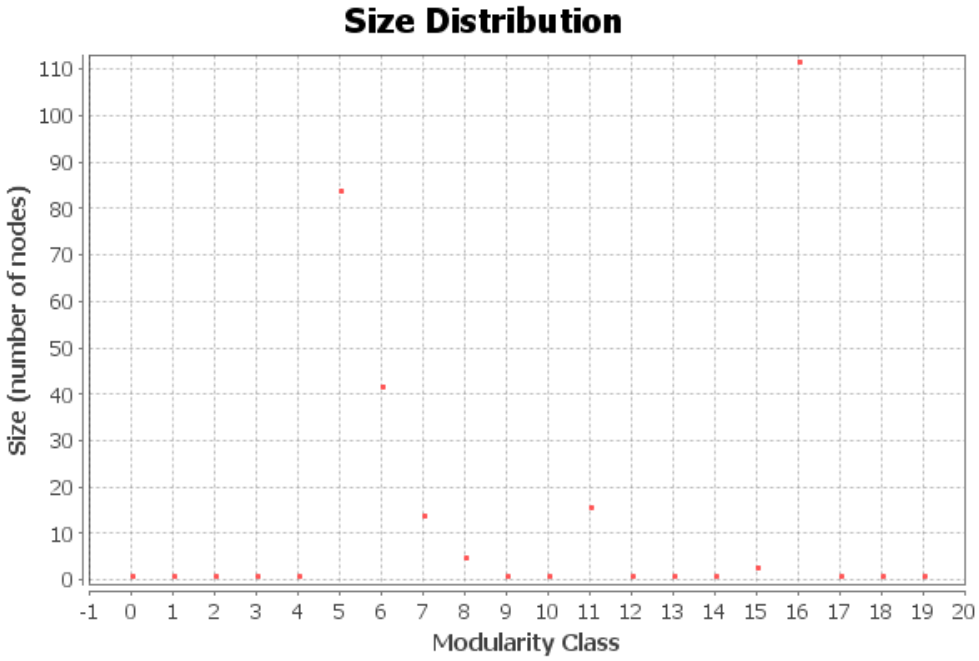
Slika 7.11 Društvena mreža korisnika u početnoj verziji

Dobivena je društvena mreža koja se sastoji od svih kontakata s kojima je autor u vezi preko njegovog korisničkog profila na Facebooku. S obzirom na brojnost kontakata i njihovih međusobnih veza, nije u potpunosti jasno tko je s kim povezan niti zbog kojeg razoga. Potrebno je provesti modularnost nad čvorovima, odnosno odrediti pripadnost čvora pojedinoj zajednici. Na prikazanoj strukturi podataka mreže je navedeno kako je svaki čvor definiran vlastitim svojstvima i vezama. Za svaki čvor je definiran identifikacijski broj, spol, broj zapisa i slično. Logično je da će osoba koja studira na FER-u imati veliki broj kolega na Facebook mreži. Vjerojatnost vezanja jednog čvora s drugim ovisi o grupama u kojima se korisnik nalazi na Facebook mreži. Na primjer ukoliko je Petar u grupi brucoša 2012/2013. vrlo je vjerojatno da je i njegov prijatelj Patrik u toj grupi te da su oni povezani u istom klasteru u mreži. Rezultat klasteriranja će biti sličan rezultatu kao i u *RStudiu*, međutim prilikom analize u ovom slučaju su implementirani i podaci koji su bili eliminirani prilikom proračuna u *RStudiu*. Nakon postupka provedbe modularizacije, rezultat je sljedeći:



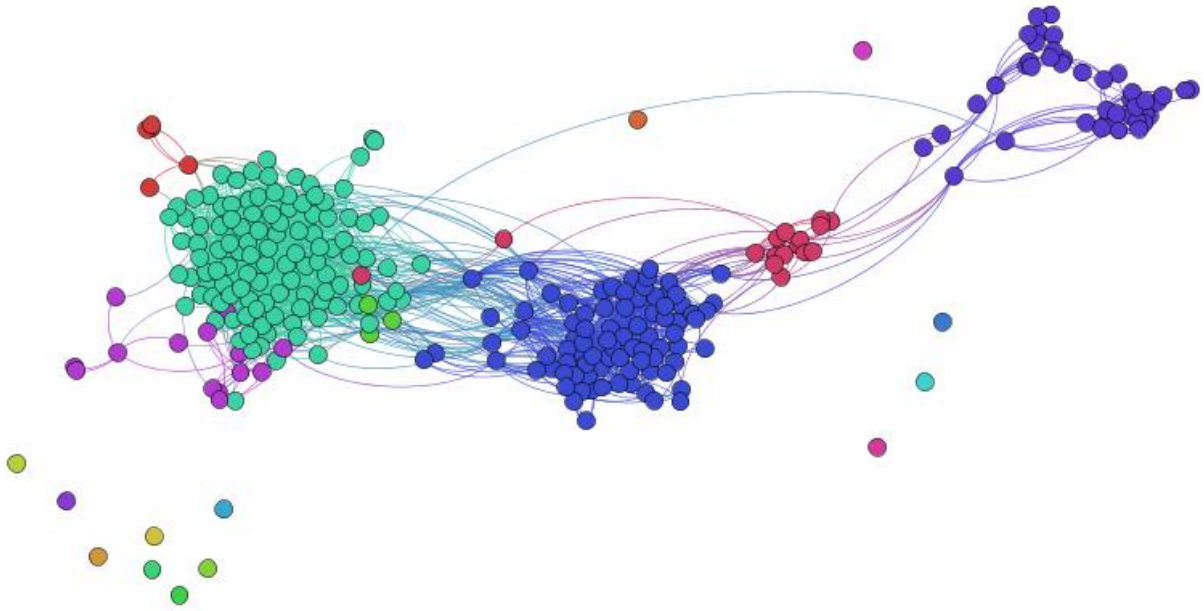
Slika 7.12 Modularizacija čvorova

Sudeći prema broju točaka, grafički je prikazan broj čvorova prema pojedinom klasteru.



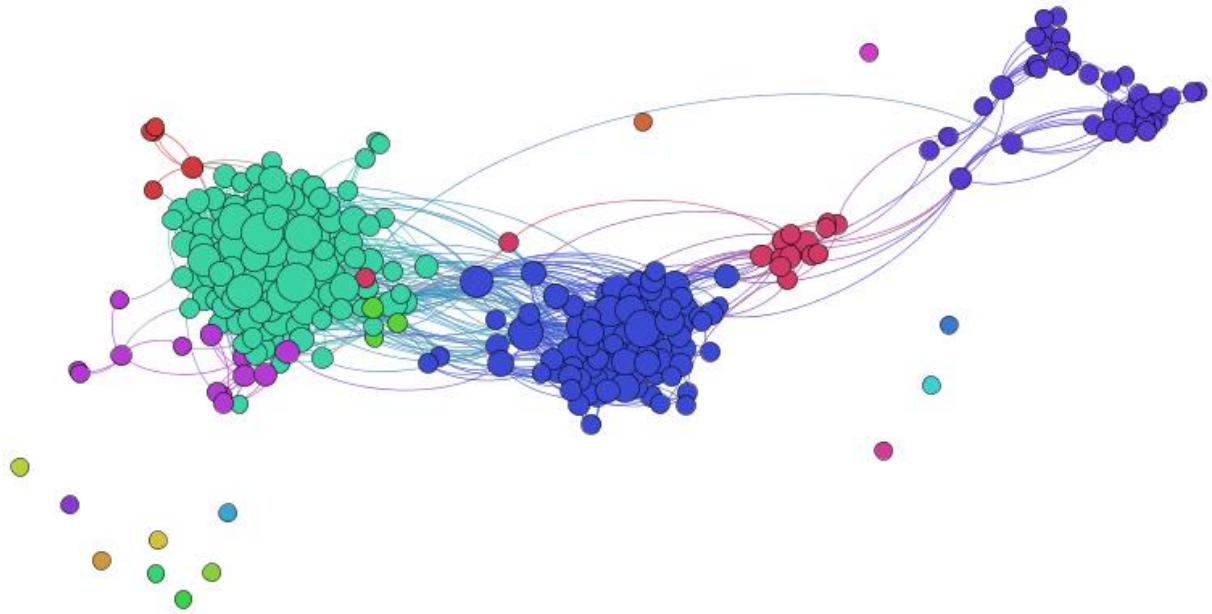
Slika 7.13 Grafički prikazana statistika za modularnost

Na slici 7.12 je bojama definirana pripadnost pojedinog čvora nekom klasteru. Za svaki čvor je definirana pripadnost, ali preglednost samog rezultata još nije u potpunosti jasna. Na slici 7.13 je prikazan kapacitet pojedinog klastera u odnosu „broj čvorova prema klasteru“. Kako bi se detaljnije prikazao svaki klaster, koristit će se model *Force Atlas 2*. *Force Atlas 2* je model koji je integriran unutar programskog okruženja *Gephi*. Model koristi moatematičke proračune slično kao i standardni algoritmi koji se mogu implementirati u *RStudiu*, međutim jedan od glavnih faktora je stupanj pojedinog čvora, na temelju kojeg vizualizira klastere. Koristeći *Force Atlas 2* provodi se raspoređivanje čvorova prema pripadnosti, pri čemu će jasno biti vidljiv pojedini klaster.



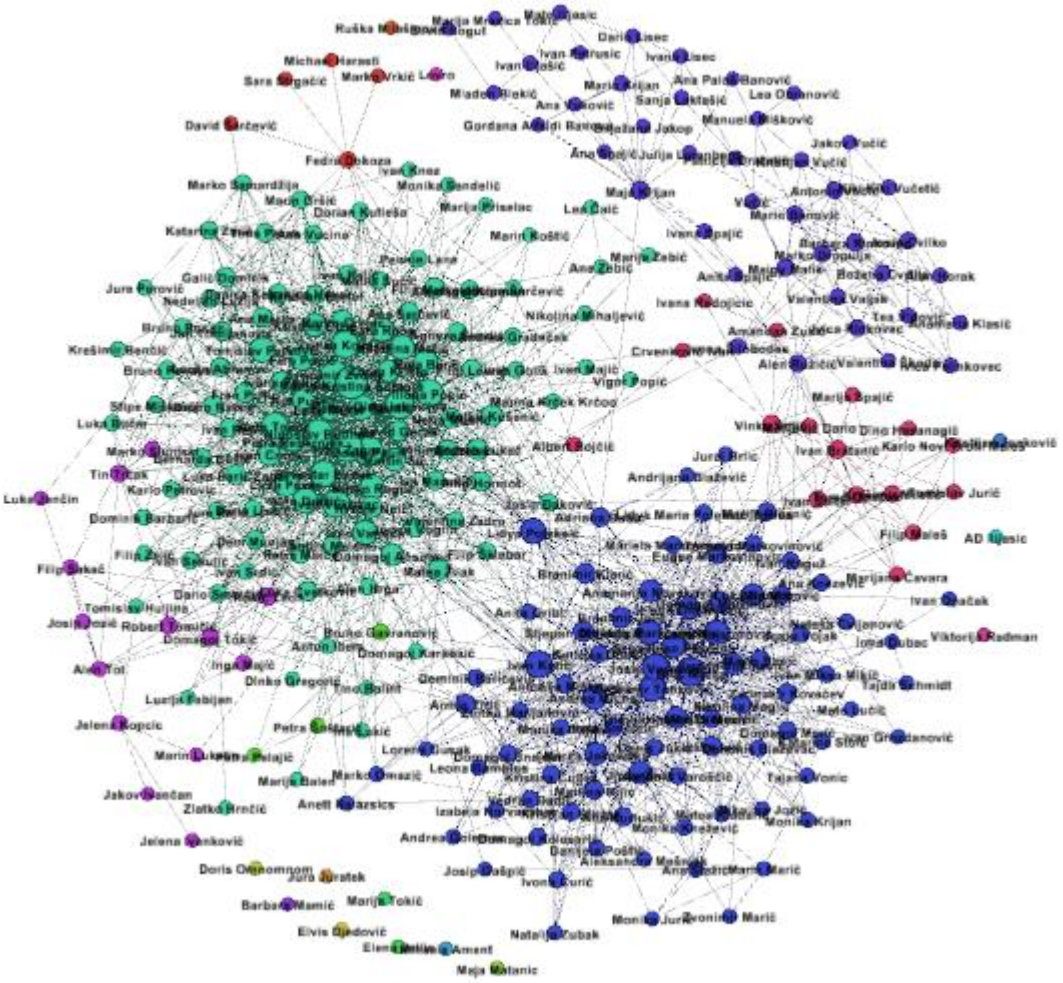
Slika 7.14 Prikaz klastera pomoću *Force Atlas 2* modela

Nakon što su klasteri vidljivi, moguće je dodatno izračunati težine pojedinog čvora, odnosno definirati za svaki čvor sa koliko je drugih čvorova povezano, te na temelju toga je dobiven rezultat prikazan na slici 7.15.



Slika 7.15 Definirana je težina za svaki čvor koja se očituje veličinom

Dobiveni rezultat može poslužiti za brojne analize. Navedeni primjer prikazuje sve klastere u društvenoj mreži autora ovog rada. Zelena skupina predstavlja sve kolege koji studiraju na istom fakultetu kao i on. Plava skupina predstavlja sve osobe koje su išle s njim u srednju školu. Ljubičasta skupina predstavlja sve članove koje su iz istog rodnog mjesta te mjesta gdje ljetuje. Čvorovi koji su udaljeni od ostalih klastera su osobe sa kojima autor rijetko kada stupa u kontakt te nema nikakvu drugu povezanost osim što su na istoj mreži. Podklasteri predstavljaju osobe koje su u suradnji s autorom na drugim projektima koji su bili i jesu aktualni u vrijeme analize podataka. Prikaz može biti drugačiji, sve ovisi o korisniku na koji način si želi vizualizirati podatke o njegovoj društvenoj mreži. Kao rezultat za prezentaciju široj skupini ljudi, ova mreža može biti predstavljena koristeći model Fruchterman-Reingold.



Slika 7.16 Prikaz društvene mreže koristeći layout Fruchterman-Reingold

Specifičnost kod modela *Frucherman-Reingold* je ta da koristi veze u drugačijem kontekstu od ostalih modela. Veze služe kao sile koje mogu biti privlačne ili odbojne. Što je veći stupanj pojedinih čvorova, veća je sila cjelokupnog klastera. Početno se definira jakost sile, prilikom čega će se čvorovi tijekom vremena dovesti u stanje konvergencije i stabilizacije. Moguće je da za neke druge podatke analiza nikada neće moći dovesti rezultat do konvergencije, što znači da glavnu ulogu imaju sami podaci. Sustav će pomicati svaki čvor dok ne dođu u adekvatni položaj, nakon čega se sustav zaustavlja.

8. Zaključak

Detekcija zajednica u društvenim mrežama je grana analize društvenih mreža koja je korisna pri analizi društvenih odnosa. Korisna je činjenica kako olakšava rad sa velikom količinom podataka. Neovisno o kojoj je društvenoj mreži riječ, na temelju metoda detekcija društvenih zajednica je lako analizirati preference pojedine zajednice. Do sada su te metode poslužile pri raznim anketama, primjerice rezultati izbora u Americi 2013. godine, generalno istraživanje glazbenih preferenci u Europi i slično. Detektiranje zajednica u društvenim mrežama je sistematičan postupak koji zahtjeva dobro poznavanje načina rada algoritama i znanje teorije grafova. Potrebno je detaljno shvaćati kakva je struktura grafa koji predstavlja društvenu mrežu, kako ne bi bilo apstraktno korisniku.

Ovim znanstvenim radom je prikazana jedan od brojnih procesa, koji mogu poslužiti kvalitetnom istraživanju društvenih mreža. U ovom primjeru je provedena analiza osoba prema školama, a općenito analiziranje može dohvatiti još veći broj podataka kao što su preferencija pojedinog glazbenog žanra među korisnicima i slično.

U budućnosti se očekuje optimiziranje postojećih algoritama za klasteriranje, koji će u manje vrijeme moći voditi proračun nad većim brojem korisnika. Do ovog trenutka su se trenutni algoritmi pokazali pouzdanima te se smatra da će povećanjem broja korisnika društvenih mreža potreba za analizom biti sve veća, a proračuni sve zahtjevniji. U daljnjem radu na ovom projektu moguće je realizirati kombinaciju nekoliko algoritama za klasteriranje u jedan, koji bi u manje vrijeme mogao obaviti cjelokupni proces. Postupak optimizacije i implementacije će biti kompleksan, međutim svaki težak put vodi do uspjeha!

9. Literatura

[1] *Mining of Massive Datasets*, 2nd Edition, Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman, pristupio 4.3.2015.

[2] Predavanje „Praktične primjene analize društvenih mreža“ iz predmeta Društvene mreže, Fakultet elektrotehnike i računarstva, Zagreb, pristupio 2.4. 2015.

[3] *Social Network Analysis*, The 3rd Edition, J. Scott, pristupio 5.4. 2015.

[4] *Social network analysis*, Serrat, O. (2010). Washington, DC, Asian Development Bank. Pristupio 5.4.2015.

[5] *Stanford Large Network Dataset Collection*, službena web-stranica fakulteta, http://snap.stanford.edu/data/index.html, pristupio 6. 4. 2015.

[6] *Gephi tutorials wiki*, službena web stranica programa *Gephi* , http://gephi.github.io/users/tutorial-layouts/ , pristupio 15. 5. 2015.

[7] Službena web stranica *RStudio*, na kojem je preuzeto programsko okruženje, http://www.rstudio.com/, pristupio 15. 5. 2015.

[8] Tutoriali za statistički alat R, http://www.statmethods.net/, pristupio 15. 5. 2015.

[9] *Using the R Statistical Computing Environment to Teach Social Statistics Courses,* Fox, John and Andersen, Robert (siječanj 2005, pristupio 17. 5. 2015.

10. Sažetak / Summary

10.1 Sažetak

Početkom društvenih mreža, broj korisnika je bio manji u odnosu na današnje razdoblje. Tijekom vremena društvene mreže su se razvijale. Danas su društvene mreže postale jedan od glavnih faktora komunikacije u svijetu. Porastom broja korisnika, samim time i podataka, bilo je potrebno provoditi analizu podataka da bi se mogli efektivnije koristiti. Radi toga je razvijena analiza društvenih mreža. U radu se proučava detekcija zajednica društvenih mreža. Detekcija zajednica društvenih mreža se koristi za analizu društvenih zajednica. Detekcija se provodi klasteriranjem, koji mogu biti hijerarhijski i na temelju sredina. Najpoznatiji je algoritam k-sredina, koji se analizirao u ovom radu na stvarnim podacima sa društvene mreže Facebook. Također su se koristili *Gephi* i *RStudio*, koji služe za analizu velikog broja podataka te njihovu vizualizaciju. Vizualizacijom algoritma je dokazana korisnos detekcije zajednica društvenih mreža u stvarnosti, a samim time je otvorena mogućnost optimiziranja postojećih načina. Detekcija zajednica društvenih mreža može uvelike pomoći prilikom analiziranja velikog broja korisnika, kako bi se društvene mreže mogle lakše prilagoditi korisnicima. Sama ta činjenica znači da će se u budućnosti društvene mreže razvijati u kvalitetne mreže te da je za to zaslužna grana matematike – analiza društvenih mreža.

10.2 Summary

At the beginning of social networks, the number of users was lower than in the present period. Over time, social networks have evolved. Today, social networks have become one of the main factors of communication in the world. As the number of users, thus the data increased, it was decided that it is necessary to conduct an analysis of data that could be used more effectively. This was the main reason for development of analysis of social networks. This paper studies the detection of community social networks. Detection of community social networks are used to benefit the user preferences and thus grouping users in simple clusters, that could be

easier to analyze. Detection is based on clustering, which can be hierarchical and based on the means. The most famous is the k-means algorithm, which is analyzed in this paper with the actual data from the social network Facebook by using the statistics tool R. Software environment Gephi and Rstudio has also been used, which primarily serves to analyze a large number of data and their visualization. Visualization of the algorithm is proven the usefulness of the implementation of the detection of community social networks in reality, and thus opened up the possibility of optimizing existing algorithms. Detection of community social networks can greatly help in analyzing a large number of users, so that social networks could more easily adapt to users. This fact alone means that in the future social networks will be developed in the high-quality network and that the credit for this success will be exact same branch of mathematics - analysis of social networks.