Detekcija zajednica u socijalnim mrežama

Seminarski rad u okviru kursa
Računarska inteligencija
Matematički fakultet

Aleksandra Nikšić, Anđelka Milovanović
mi16072@alas.matf.bg.ac.rs, mi15145@alas.matf.bg.ac.rs

15. april 2020.

Sažetak

Ovaj rad se bavi metodama detekcije zajednica u socijalnim mrežama. Istraživanje je odrađeno sa Girvan-Newman algoritmom sa običnom i gustinskom modularnošću, kao i sa Meme-Net algoritmom i njegovom modifikacijom sa heuristikom simuliranog kaljenja umesto lokalne pretrage. Zajednice su bile poznati skupovi podataka: Karate, Delfini i Jazz.

Sadržaj

1 Uvod

Otkrivanje zajednica [2] je važno istraživačko polje u analizi socijalnih mreža gde se bavimo uočavanjem strukture mreže. Struktura zajednice je ključna topološka karakteristika kompleksnih mreža. Savremena nauka o mrežama donela je značajan napredak našem razumevanju ovih složenih sistema.

Problem predstavlja uočavanje jače i slabije povezanih delova mreže na osnovu posmatranja isključivo eksplicitnih veza između čvorova. Grupe čvorova unutar kojih su mrežne veze guste predstavljaju zajednice, dok su veze između takvih grupa niže gustine. Ovakve zajednice mogu se smatrati prilično nezavisnim komponentama grafa, igrajući sličnu ulogu poput tkiva ili organa u ljudskom telu. Sposobnost pronalaženja i analiziranja takvih grupa može značajno doprineti razumevanju i vizualizaciji strukture mreža.

Detekcija zajednica nalazi veliku primenu u naukama poput sociologije, biologije ili računarstva, disciplinama u kojima su sistemi često predstavljeni kao grafovi. Primenom algritama detekcije na realne mreže možemo zaključiti osobine i veze između čvorova, koje nisu dostupne iz direktnog posmatranja grafa.

2 Algoritmi

Jednostavan način identifikovanja zajednica na grafu je detektovanje grana koje povezuju čvorove različitih zajednica i njihovo uklanjanje, kako bi se klasteri razdvojili [6]. Sa druge strane, zajednice se mogu tražiti genetskim algoritmom gde su jedinke predtavljene kao nizovi čvorova sa određenim pripadnostima. U naredne dve sekcije biće predstavljeni ovi algoritmi, jer su oni korišćeni za analizu detekcija zajednica.

2.1 Girvan-Newman

Algoritam koji je obeležio početak nove ere na polju detekcije zajednica predložen je od strane naučnika Girvana i Newmana [6]. U njemu se grane biraju u odnosu na vrednost njihovog centraliteta (eng. *edge centrality*), koja procenjuje važnost grane. Koraci algoritma implementirani od strane Jahanbakhsh [4] su sledeći:

1. postavi se najbolja modularnost na BestQ=0
2. učita se graf i izračunaju se njegove komponente
3. izračuna se edge-betweenness za sve ivice i obrišu se sve one sa maksimalnom vrednošću (one su most između zajednica)
4. izračuna se novi broj komponenti grafa
5. ako je novi broj komponenti $\leq$ od početnog onda se ponavlja korak 3
6. izračuna se modularnost i sačuva se u Q
7. ako je Q>BestQ onda se ažurira najbolja modularnost i sačuva se ta podela grafa kao najbolja u BestComps
8. ako nema više ivica u grafu vraća se BestComps, u suprotnom se ponavlja proces od koraka 3 na dalje

Fokus algoritma je na koceptu međusobnosti (eng. *betweenness*), promenljivoj koja izražava frekvenciju učešća grane u procesu. Složenost izračunavanja ove promenljive je O(mn), a na retkim grafovima O(n^2). Razmatrane su tri različite vrste, a ona koja je eksperimentalno dala najbolje rezultate je EB (eng. *edge betweenness*). Ona govori koliko najkraćih putanja između svih parova čvorova na grafu u svom skupu grana sadrži baš posmatranu granu. Intuitivno je da grane unutar zajednice imaju veliku vrednost EB, zato što će dosta najkraćih putanja koje povezuju čvorove različitih zajednica preći preko njih.

Ipak, algoritam je prilično spor i primenljiv na retke grafove sa do 10000 čvorova. Originalni algoritam nije imao proceduru za biranje najbolje particije, ali se usavršavanjem došlo se načina odabira. Particija koja će biti odabrana ima najveću vrednost modularnosti (eng. *modularity*), kriterijuma koji se od tada frekventno koristi.

Girvan Newman modularnost Q, prvobitno je uvedena radi definisanja kriterijuma zaustavljanja za algoritam Girvana i Newmana, brzo je postala suštinski element mnogih metoda za detekciju zajednica. Ona je dosta korišćena funkcija kvaliteta.

Pretpostavimo da visoke vrednosti modularnosti ukazuju dobre particije (nije generalno tačno). Dakle, particija koja odgovara njenoj maksimalnoj vrednosti na datom grafu treba da bude najbolja, ili barem vrlo dobra. To je glavna motivacija za maksimizaciju modularnosti. Iscrpna optimizacija Q-a je nemoguća zbog ogromnog broja načina na koje je moguće particionisati graf. Osim toga, stvarni maksimum je van domašaja, jer je nedavno dokazano da je optimizacija modularnosti NP-kompletan problem. Međutim, trenutno postoji nekoliko algoritama koji mogu pronaći poprilično dobre aproksimacije maksimuma modularnosti u razumnom vremenu. Modularnost je definisana na sledeći način:

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - P_{ij}) \delta(C_i, C_j) \tag{1}$$

gde suma obuhvata sve parove čvorova, A je matrica susedstva, m je ukupan broj grana grafa, a P_{ij} predstavlja očekivani broj grana između čvorova i i j u originalnom grafu i δ funkcija ima vrednost 1 ako su čvorovi i i j u istoj zajednici ($C_i = C_j$), a 0 inače.

Za datu particiju grafa $\omega = V_1, V_2, ..., V_m$, gde je V_i skup čvorova podgrafa G_i za $i = 1, ..., m$, modularnost zasnovana na gustini (poznata kao D-vrednost) je definisana na sledeći način [5]:

$$D = \sum_{i=1}^{m} \frac{L(V_i, V_i) - L(V_i, \overline{V}_i)}{|V_i|} \tag{2}$$

gde su skupovi čvorova V_1 i V_2 disjunktni podskupovi skupa čvorova grafa V. Definišemo $L(V_1, V_2) = \sum_{i \in V_1, j \in V_2} A_{ij}$, $L(V_1, V_1) = \sum_{i \in V_1, j \in V_1} A_{ij}$ i $L(V_1, \overline{V}_1) = \sum_{i \in V_1, j \in \overline{V}_1} A_{ij}$, gde je $\overline{V}_1 = V - V_1$. U ovoj jednačini, svaki sabirak označava odnos između razlike unutrašnjeg i spoljašnjeg stepena podgrafa G_i i veličine podgarafa. Što je vrednost D veća, particija je bolja. Dakle, problem detekcije zajednice se može posmatrati kao problem pronalaženja particije mreže takve da je modularnost zasnovana na gustini maksimizirana.

Ova funkcija kvaliteta se može i generalizovati [3]:

$$D_\lambda = \sum_{i=1}^{m} \frac{2\lambda L(V_i, V_i) - 2(1-\lambda) L(V_i, \overline{V}_i)}{|V_i|} \tag{3}$$

Variranjem λ možemo posmatrati mrežu u različitim rezolucijama. Parametar λ može uzeti vrednost u opsegu od 0 do 1. Za veće vrednosti λ ovaj metod ima tendenciju da zajednicu razbija na više manjih zajednica. Za manje vrednosti λ dobija se podela na veće zajednice.

2.2 Memetski algoritam

Za detekciju zajednica u okviru memetskog algoritma [3] optimizovaćemo modularnost zasnovanu na gustini. Razmatraćemo i podesivi parametar λ koji nam daje mogućnost istraživanja mreže u različitim rezolucijama. Ovaj algoritam je originalno simbioza genetskog algoritma i strategije lokalne pretrage. U ovom radu heuristiku lokalne pretrage zamenićemo tehnikom simuliranog kaljenja.

U standardnom genetskom algoritmu, populacija stringova (hromozoma), koja enkodira kandidate (jedinke) za rešenje optimizacionog problema, evoluira ka boljim rešenjima. Evolucija najčešće kreće iz nasumično izabrane populacije. U svakoj generaciji, izračunava se fitnes svake jedinke, bira se određen broj jedinki koji će učestvovati u modifikaciji (operatorima ukrštanja i mutacije) da bi oformili novu populaciju. Nakon nekoliko generacija, samo ona rešenja sa velikom vrednošću fitnesa će preživeti. U radu Gonga i drugih naučnika [3], modularnost je fitnes funkcija, a particije su hromozomi. Evolutivni algoritmi koji rekombinaciju viskokvalitetnih rešenja presecaju pojedinačnim optimizacijama nazvani su memetski algoritmi. Metode su inspirisane modelima prirodnih sistema koji kombinuju evolucijsku adaptaciju populacije sa individualnim učenjem tokom životnog veka njenih članova. Više o tome kako se generiše inicijalna populacija i na koji način se vrše operacije ukrštanja i mutacije jedinki pročitati u radu [3], a implementacioni detalji se mogu pogledati na adresi: `https://github.com/newxd/community-detection`.

Simulirano kaljenje iskoristićemo kako bismo poboljšali pojedinačna rešenja. Pri ažuriranju najbolje jedinke pretraživaćemo njenu okolinu i analizirati funkcije cilja datih rešenja. Ukoliko naiđemo na rešenje sa boljom vrednošću funkcije cilja, ažuriraćemo najbolje rešenje pronađenim. Ukoliko naiđemo na rešenje sa lošijom vrednošću funkcije cilja, poredićemo vrednosti unapred definisane funkcije p i proizvoljno izabrane vrednosti q iz intervala (0,1). Ako je $p > q$, ažuriraćemo najbolje rešenje pronađenim.

3 Rezultati

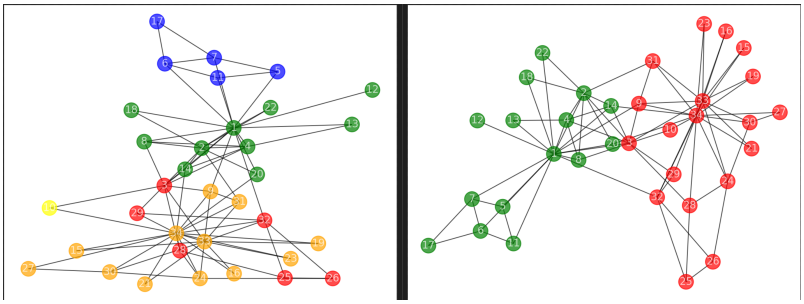
U ovoj sekciji biće prestavljeni rezultati dobijeni nad podacima skupova Karate, Delfin i Jazz. Nad svakim skupom eksperimentisano je sa dve verzije Girvan-Newman algoritma (obična modularnost [2] i modularnost sa gustinom [5]), kao i sa verzijama Memetskog algoritma sa podrazumevanom lokalnom pretragom (Meme-Net [3]) i sa dodatom heuristikom simuliranog kaljenja koje je isprobano sa ciljem eksperimentisanja.

3.1 Karate skup

Karate skup (eng. Zachary' s karate club) se sastoji od 34 čvora i 78 ivica. Konstruisao ga je Zachary, posmatrajući 34 člana karate kluba tokom perioda od 2 godine. Nakon svađe između šefa kluba i instruktora, zajednica se podelila na 2 dela, jer je instruktor napravio svoj klub i odveo oko polovine članova [3].

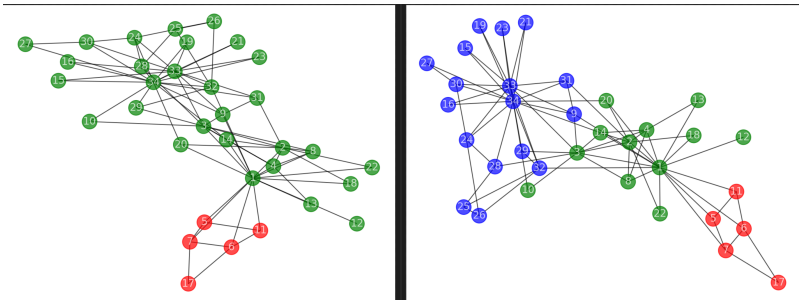
Gong et al. su u svom radu osmislili Meme-Net algoritam, koji je rekonstruisan u ovom radu. Dodatak tom algoritmu jeste simulirano kaljenje koje je isprobano u delovima u kojima je bilo predviđeno da se koristi lokalna pretraga. Pored eksperimentisanja sa simuliranim kaljenjem u Meme-Net algoritmu, isprobana je i verzija GN algoritma gde se za funkciju modularnosti koristi gustinska modularnost.

Na slici 1 se jasno može videti da je podela sa gustinskom modularnošću bolje podelila graf, što je potvrdila i NMI (eng. Normalized Mutual Information) veličina, koja je za običan GN bila 0.58, dok je za izmenjenu modularnost iznosila NMI: 0.84. Menjanjem parametara ovo se nije moglo unaprediti dalje. Na slici 2 su predstavljeni rezultati dobijeni pokretanjem



Slika 1: Sa leve strane slike je podrazumevani GN, dok je sa desne strane GN sa gustinskom modularnošću sa parametrom λ=0.3

Meme-Net algoritma za 20 generacija i λ=0.5. Sa leve strane je Meme-Net, a sa desne strane izmenjena verzija sa simuliranim kaljenjem od 500 iteracija. NMI vrednosti su bile redom: 0.22 i 0.7. Kada se λ parametar

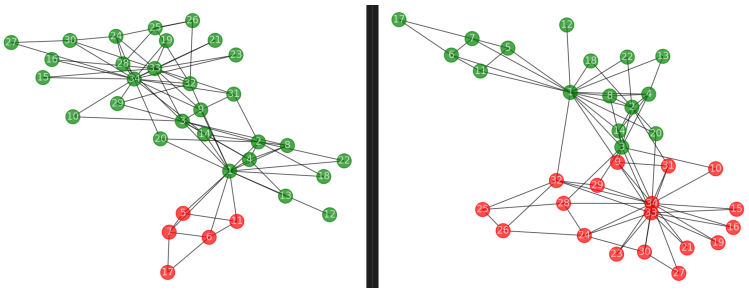


Slika 2: Sa leve strane slike je Meme-Net, dok je sa desne strane Meme-Net sa simuliranim kaljenjem (oba za 20 generacija) i λ=0.5

postavi na 0.3 vrednost (predloženu u [3]), dobija se NMI=1 sa simuliranim kaljenjem za 20 generacija, dok običan Meme-Net ima značajno manju NMI nakon toliko generacija. Rezultati su predstavljeni na slici 3.

3.2 Delfin skup

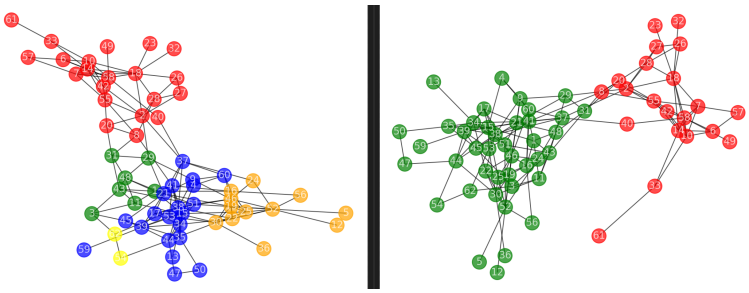
Skup delfina sastoji se od 62 delfina i 159 veza između njih. Delfini su sa Novog Zelanda i podaci o njima su prikupljani tokom 7 godina. Veza tj.



Slika 3: Sa leve strane slike je Meme-Net, dok je sa desne strane Meme-Net sa simuliranim kaljenjem (oba za 20 generacija) i λ=0.3

ivica između 2 delfina je ostvarena na osnovu njihovog statistički čestog druženja. Prirodno se mreža deli u 2 velike grupe [3].

Ono što se može zaključiti iz GN algoritma je kao i kod Karate skupa, da se izmenom modularnosti na gustinsku modularnost, može postići veći NMI, ali ne može dostići vrednost 1. Rezultati su prikazani na slici 4. Što se tiče memetskog algoritma, nakon 50 generacija NMI dostiže vrednost 1 sa simuliranim kaljenjem, dok lokalna pretraga dostiže vrednost 0.48. Rezultati su prikazani na slici 5.

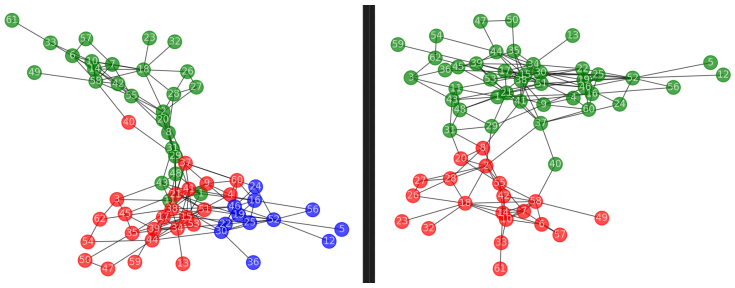


Slika 4: Sa leve strane slike je podrazumevani GN, dok je sa desne strane GN sa gustinskom modularnošću sa parametrom λ=0.3.

3.3 Jazz skup

Jazz skup predstavlja mrežu muzičara. Ukupan broj čvorova je 198 i svaki čvor predstavlja jednog muzičara. Broj ivica je 5484 i ivica između muzičara postoji ukoliko su 2 muzičara bili članovi istog benda. Podaci su sakupljeni tokom 2003. godine [1].

Izveštaj sa informacijama za ovaj skup biće predstavljen na odbrani seminarskog rada.



Slika 5: Sa leve strane slike je Meme-Net, dok je sa desne strane Meme-Net sa simuliranim kaljenjem (oba za 50 generacija) i λ=0.3

4 Zaključak

Tokom istraživanja oblasti detekcije zajednica u socijalnim mrežama, zaključeno je da se u zavisnosti od problema različiti algoritmi ponašaju bolje ili lošije. Naš rad je ograničen na istraživanje Girvan-Newman algoritma, kao i memetskog algoritma.

Rezultati su pokazali da se izmenom klasične modularnosti sa gustinskom modularnošću kod algoritma Girvan-Newman dobijaju značajno bolje particije grafa. To je utvrđeno poređenjem sa stvarnom particijom skupova Karate i Delfin, koristeći NMI meru.

Što se tiče izmene u Meme-Net algoritmu u kome smo umesto lokalne pretrage primenili heuristiku simuliranog kaljenja, ne može se sa sigurnošću reći da uvek radi brže i bolje. Međutim, za ispitane slučajeve konvergencija je bila brža i NMI je dostizao vrednost 1.

Dalji rad bi mogao da se fokusira na istraživanje većih zajednica ovim algoritmima.

Literatura

[1] Jazz musicians network dataset – KONECT, September 2016.

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

[3] Maoguo Gong, Bao Fu, Licheng Jiao, and Haifeng Du. Memetic algorithm for community detection in networks. *Physical Review E*, 84(5):056101, 2011.

[4] Kazem Jahanbakhsh. Community Detection in Social Networks, 2010. on-line at: http://www.kazemjahanbakhsh.com/codes/cmty.html.

[5] Zhenping Li, Shihua Zhang, Rui-Sheng Wang, Xiang-Sun Zhang, and Luonan Chen. Quantitative function for community detection. *Physical review E*, 77(3):036109, 2008.

[6] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

A Dodatak

GitHub repozitorijum seminarskog rada se može naći na adresi: https://github.com/mandja96/RI-detekcija_zajednica.