Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Bine Iljaš

Gručenje s pomočjo druge najmanjše lastne vrednosti Laplaceove matrike

diplomsko delo

univerzitetni študijski program
prve stopnje
računalništvo in informatika

Mentorica: izr. prof. dr. Polona Oblak
Somentorica: as. dr. Aleksandra Franc

Ljubljana, 2019

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Tematika naloge:

V delu preučite lastnosti Laplaceove matrike grafa ter njenih lastnih vrednosti in vektorjev. Posebej se posvetite Fiedlerjevemu lastnemu vektorju druge najmanjše lastne vrednosti. Implementirajte algoritem, ki s pomočjo komponent Fiedlerjevega vektorja naredi gruče vozlišč grafa, ter ga primerjajte z ostalimi metodami gručenja.

Zahvaljujem se svoji mentorici izr. prof. dr. Poloni Oblak ter somentorici as. dr. Aleksandri Franc za pomoč in vodenje pri pisanju diplomskega dela. Zahvalil bi se tudi staršem, bratu in prijateljem, ki so me pri delu spodbujali.

Kazalo

Seznam uporabljenih kratic

kratica	angleško	slovensko
FRI	Faculty of computer and information science	Fakulteta za računalništvo in informatiko
SICRIS	Slovenian Current Research Information System	Informacijski sistem o raziskovalni dejavnosti v Sloveniji

Povzetek

Naslov: Gručenje s pomočjo druge najmanjše lastne vrednosti Laplaceove matrike

Avtor: Bine Iljaš

V diplomski nalogi opišemo osnovne koncepte grafov ter njim pripadajočih matrik. Posebej preučimo Laplaceovo matriko ter lastnosti lastnega vektorja, ki pripada njeni drugi najmanjši lastni vrednosti. S pomočjo Fiedlerjevega vektorja razdelimo graf v skupine. Predstavimo metodo voditeljev, metodo deljenja grafa po predznakih ter metodo maksimiziranja modularnosti. Implementiramo algoritem, ki z metodo voditeljev ali rekurzivno metodo delitve grafa po predznakih razdeli vozlišča danega grafa na podskupine. Izvedemo nekaj poskusov na umetnih, nato pa še na realnih podatkih. Rezultate v delu predstavimo in jih primerjamo z metodo maksimiziranja modularnosti. Z Randovim indeksom ocenimo uspešnost delitve grafa. Ugotovimo, da na umetnih podatkih vse metode delujejo približno enako dobro, medtem ko na realnih podatkih metoda maksimiziranja modularnosti deluje nekoliko bolje od preostalih dveh.

Ključne besede: grafi, Laplaceova matrika, Fiedlerjev vektor, gručenje, metoda voditeljev, modularnost.

Abstract

Title: Clustering with the Second Smallest Eigenvalue of the Laplacian Matrix

Author: Bine Iljaš

In this work, the basic concepts of graphs and their corresponding matrices are discussed, as well as the Laplacian matrix and the properties of the eigenvector. The results demonstrate that the components of the Fiedler vector (vector corresponding to the second smallest eigenvalue of the Laplacian matrix), divide the graph into two groups. The following clustering methods are outlined: k-means, the spectral partitioning method and the modularity maximization method. An algorithm was implemented that uses k-means, or the recursive method, to divide the graph into subgroups. The experiment was carried out using artificial data as well as real data. The results are presented and compared with the method of maximizing the modularity. The Rand index is used to evaluate the success of the clustering. It is discovered that when using artificial data, all methods give comparable results, however using the method of maximizing the modularity gives slightly better results than the other two when using real data.

Keywords: graphs, Laplacian matrix, Fiedler vector, clustering, k-means, modularity.

Poglavje 1

Uvod

1.1 Motivacija

Danes živimo v časih, ko se na različnih področjih srečujemo z veliko količino informacij. V digitalni dobi nam preobilica podatkov onemogoča, da bi podatke pregledali ročno, saj bi bilo to časovno prezahtevno.

Nekatere podatke lahko enostavno in na razumljiv način prikažemo z grafom. Računalniku lahko graf predstavimo v obliki matrike ali pa zapisa vozlišč in povezav med njimi. Ker obstaja veliko knjižnic za delo z matrikami, je včasih smiselno graf predstaviti kot matriko in za operacije z njimi uporabiti že implementirane algoritme.

Podatki, ki se lepo predstavijo z grafi, so lahko na primer prijateljstva na družabnih omrežjih. Še en primer uporabe grafa je mreža računalnikov, ki so povezani v omrežje. Računalniki povezani v notranja omrežja ustvarjajo več med seboj povezanih skupin.

V tej diplomski nalogi bomo v podatkih iskali skupine tako, da bomo pripadajoče grafe razdelili na več delov, ki so med seboj povezani. Na probleme naletimo v vozliščih, ki bi lahko smiselno pripadala dvema ali večim različnim skupinam. V takih primerih se je potrebno odločiti, kateri skupini dodeliti omenjeno vozlišče. Za ta namen bomo implementirali dve metodi, in sicer metodo voditeljev, ki jo bomo izvedli na komponentah Fiedlerjevega

vektorja, ter metodo, ki rekurzivno deli graf na dve skupini glede na predznake komponent Fiedlerjevega vektorja. Napisani metodi bomo primerjali z metodo gručenja, ki maksimizira modularnost grafa. Podobne tematike, predvsem Fiedlerjev vektor in delitev grafov, obravnava kar nekaj člankov in znanstvenih del. Brian Slininger v članku Fiedler's Theory of Spectral Graph Partitioning [14] opiše Laplaceovo matriko in njene lastne vrednosti ter pripadajoče lastne vektorje. Na primeru nazorno prikaže delitev grafa na dve skupini glede na predznake komponent Fiedlerjevega vektorja. Santo Fortunato v članku Community detection in graphs [3] celotno četrto poglavje nameni opisu metod iskanja skupin na grafih. V knjigi Marka Newmana Networks: An Introduction [8] pa je v poglavju 11.5 podrobno opisana metoda spektralne particije.

1.2 Cilj

Za prvi cilj diplomske naloge smo si zadali preučiti Laplaceovo matriko in njene lastne vektorje. Opisali bomo različne metode gručenja ter iskanja gosto povezanih skupin. V delu bomo predstavili matematične koncepte Laplaceove matrike ter njenih lastnih vrednosti in vektorjev. V drugem delu diplomske naloge jih bomo uporabili pri implementaciji algoritma za iskanje skupin v grafih. Graf bomo s pomočjo lastnega vektorja Laplaceove matrike razdelili na skupine z metodo voditeljev ter z metodo delitve po predznakih. Po končani implementaciji bomo delovanje napisanega algoritma primerjali z obstoječo metodo iskanja skupin v grafih. Za primerjavo bomo uporabili metodo *FindGraphCommunities*, ki je vgrajena v Wolfram Mathematici. Za ocenitev podobnosti bomo uporabili Randov indeks in z njim izmerili podobnost ustvarjenih skupin.

1.3 Struktura

V drugem poglavju diplomskega dela bomo predstavili grafe ter Laplaceovo matriko in njene lastne vektorje. V tretjem poglavju bomo predstavili uporabljene metode gručenja, v četrtem in petem poglavju sledita opisa implementacije rešitve in podatkov, ki smo jih uporabili. Opisali bomo tudi izvedene teste ter predstavili dobljene rezultate. V zaključku sledi še povzetek celotne naloge in rezultati.

Poglavje 2

Graf in Laplaceova matrika

2.1 Graf

Graf je struktura, sestavljena iz množice objektov in množice povezav. Objekte v grafu predstavimo z vozlišči, povezave pa z (neurejenimi) pari vozlišč.

Definicija 2.1.1 *Graf je urejena množica $G = (V, E)$, kjer je V neprazna množica vozlišč, E pa množica povezav, to je množica parov vozlišč.*

Včasih grafu $G = (V, E)$, pridružimo še funkcijo $w : E \to \mathbb{R}$, ki vsaki povezavi grafa pridruži realno število. Tedaj govorimo o uteženem grafu, vrednosti $w_{i,j}$ pa imenujemo uteži. Več o grafih si bralec lahko prebere v devetem poglavju knjige Diskretne strukture [2] ali denimo v članku Graf – podatkovna struktura [11].

2.2 Grafom pripadajoče matrike

2.2.1 Matrika sosednosti

Vsak graf lahko predstavimo z matriko sosednosti. Matrika sosednosti neusmerjenega grafa je simetrična.

Elementi matrike sosednosti $A(G) = [a_{i,j}], i, j = 1, ..., |V|$, so za neutežen graf $G = (V, E)$ definirani kot

$$a_{i,j} = \begin{cases} 1, & \{i,j\} \in E, \\ 0, & \{i,j\} \notin E. \end{cases}$$

Če sta vozlišči i in j sosednji, sta elementa na križišču i-te vrstice in j-tega stolpca ter j-te vrstice in i-tega stolpca enaka 1.

2.2.2 Matrika stopenj

Druga pomembna matrika je matrika stopenj $D(G) = [d_{i,j}], i, j = 1, ..., |V|$. Matrika stopenj je diagonalna matrika. Stopnja vozlišča i, $\mathrm{Deg}(i)$, je enaka številu povezav, ki izhajajo iz tega vozlišča. Definicija elementov matrike stopenj je:

$$d_{i,j} = \begin{cases} \mathrm{Deg}(i), & i = j, \\ 0, & i \neq j, \end{cases}$$

kjer $\mathrm{Deg}(i)$ označuje stopnjo vozlišča i.

2.2.3 Laplaceova matrika

Vsakemu grafu je mogoče prirediti Laplaceovo matriko $L = [l_{i,j}], i, j = 1, ..., |V|$. Če sta vozlišči i in j povezani, potem v matriki v i-to vrstico in j-ti stolpec ter v j-to vrstico in i-ti stolpec zapišemo vrednost -1, na diagonalo matrike pa zapišemo stopnje vozlišč:

$$l_{i,j} = \begin{cases} \mathrm{Deg}(i), & i = j, \\ -1, & i \neq j \text{ in } i \text{ soseden z } j, \\ 0, & \text{v ostalih primerih.} \end{cases}$$

Laplaceova matrika je razlika matrike stopenj in matrike sosednosti, ki pripadata istemu grafu, $L(G) = D(G) - A(G)$.

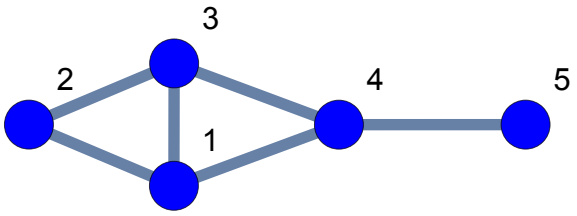
Podobno lahko definiramo elemente Laplaceove matrike za utežene grafe:

$$l_{i,j} = \begin{cases} \sum_{i,j \in E} w_{i,j}, & i = j, \\ -w_{i,j}, & i \neq j \text{ in } i \text{ soseden z } j, \\ 0, & \text{v ostalih primerih.} \end{cases}$$

Pri tem smo z $w_{i,j}$ označili vrednost uteži.

Na sliki 2.1 je prikazan graf, ki mu pripada Laplaceova matrika

$$M = \begin{bmatrix} 3 & -1 & -1 & -1 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 \\ -1 & 0 & -1 & 3 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix}.$$



Slika 2.1: Graf, ki mu pripada Laplaceova matrika M.

2.3 Lastna vrednost in lastni vektor

Lastna vrednost $n \times n$ matrike A je skalar λ, pri katerem je za neničelni vektor x izpolnjena enakost $Ax = \lambda x$. Takšen vektor x se imenuje *lastni vektor*.

Lastne vrednosti matrike A so rešitve enačbe $\det(A - \lambda I) = 0$, torej ničle polinoma stopnje n.

2.4 Lastni vektor najmanjše lastne vrednosti

Naj bo G graf in L pripadajoča Laplaceova matrika. Ker vemo, da je L simetrična in pozitivno semidefinitna matrika [9], so vse njene lastne vrednosti nenegativna realna števila. Pozitivno semidefinitna matrika je matrika L, za katero velja, da lahko poiščemo tako matriko B, da $L = BB^T$. Matrika $B[n \times m]$ je incidenčna matrika grafa z n vozlišči in m povezavami. Stolpci matrike predstavljajo povezave, vrstice pa vozlišča grafa. To pomeni, da imata elementa v k-tem stolpcu, ki pripada povezavi e_k, ki povezuje vozlišči v_i in v_j v incedenčni matriki v i-ti vrstici vrednost 1 in v j-ti vrstici vrednost -1. Ostali elementi incidenčne matrike so enaki 0.

Ker je vsota elementov vsake vrstice Laplaceove matrike enaka 0, so vrstice matrike med seboj linearno odvisne, to pomeni, da matrika nima polnega ranga in ima zato vsaj eno lastno vrednost enako 0. Lastnim vrednostim take matrike lahko izberemo n pripadajočih lastnih vektorjev tako, da so paroma pravokotni. Lastne vrednosti matrike L uredimo po velikosti tako, da je $0 = \lambda_0 \leq \lambda_1 \leq ... \leq \lambda_{n-1}$. Lastni vektor, ki pripada λ_i, označimo z v_i. Če je $\lambda_0 \neq \lambda_1$, je graf povezan. Če ima graf k komponent, je $\lambda_0 = \lambda_1 = ... = \lambda_{k-1} = 0$, pripadajoče lastne vektorje pa lahko izberemo tako, da ima vsak enice na mestih, ki ustrezajo vozliščem ene od komponent, na ostalih mestih pa ničle [10]. Na primer, če ima graf dve komponenti, vsaka od njiju vsebuje tri vozlišča, lahko lastna vektorja za $\lambda_0 = \lambda_1 = 0$ izberemo tako, da je $v_0 = [1, 1, 1, 0, 0, 0]^T$ in $v_1 = [0, 0, 0, 1, 1, 1]^T$, če ena komponenta vsebuje vozlišča 1, 2 in 3, druga pa vozlišča 4, 5 in 6. Če so vsa vozlišča grafa s 6 vozlišči v isti komponenti, je $\lambda_0 \neq \lambda_1$ in $v_0 = [1, 1, 1, 1, 1, 1]^T$ je lastni vektor za λ_0. Tega tu ne bomo dokazovali, bralec si lahko dokaz prebere v članku A Short Tutorial on Graph Laplacians, Laplacian Embedding, and Spectral Clustering [4].

2.5 Lastni vektor druge najmanjše lastne vrednosti

Prva lastna vrednost grafa z eno samo komponento, ki je večja od 0 (λ_1), je lastna vrednost, ki pripada vektorju, ki ga uporabimo za deljenje grafa. Lastni vektor druge najmanjše lastne vrednosti λ_1 je poznan tudi kot Fiedlerjev vektor. Druga najmanjša lastna vrednost se imenuje tudi algebraična povezanost.

Naj K_n označuje poln graf z n vozlišči, kjer je vsako vozlišče povezano z vsemi ostalimi. Algebraična povezanost grafa K_{100} je 100. Ciklični graf C_n je povezan graf, kjer ima vsako vozlišče natanko dva soseda. Algebraična povezanost grafa C_{100} je enaka 0.00395. Ker je graf K_{100} veliko bolj gosto povezan kot graf C_{100}, je algebraična povezanost grafa K_{100} večja kot algebraična povezanost grafa C_{100}.

Fiedler [14] je opazil, da lastni vektor druge najmanjše lastne vrednosti graf razdeli na dva dela tako, da je število povezav med njima čim manjše. Komponente Fiedlerjevega vektorja je razdelil v dve skupini in sicer na tiste s pozitivnim in tiste z negativnim predznakom.

Za graf $G = (V, E)$ in grafu pripadajočo Laplaceovo matriko L ter vsak $x \in \mathbb{R}^n$ velja

$$x^T L x = \sum_{\{i,j\} \in E} (x_i - x_j)^2.$$

To se prepričamo s pomočjo naslednjega izračuna:

$$\begin{aligned} x^T Lx &= \sum_{i,j=1}^{n} l_{i,j} x_i x_j \\ &= \sum_{i,j=1}^{n} (d_{i,j} - a_{i,j}) x_i x_j \\ &= \sum_{i=1}^{n} d_{i,i} x_i^2 - \sum_{\{i,j\}\in E} 2x_i x_j \\ &= \sum_{\{i,j\}\in E} (x_i^2 + x_j^2 - 2x_i x_j) \\ &= \sum_{\{i,j\}\in E} (x_i - x_j)^2. \end{aligned}$$

Želimo torej najti tak lasten vektor x, ki ustreza enakosti

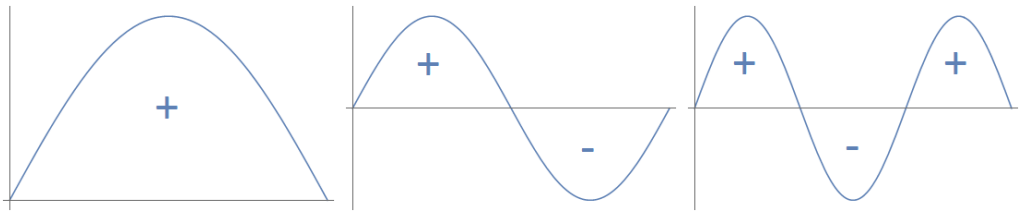
$$\lambda_1 = \min_x \sum_{\{i,j\}\in E} (x_i - x_j)^2.$$

Lastni vrednosti λ_1 pripada neničelni lastni vektor x, za katerega velja zgornja enakost in je pravokoten na vektor, ki pripada λ_0. Ker skupine ustvarimo glede na predznake komponent v vektorju x, opazimo, da povezave med vozlišči v isti skupini ne zvišujejo vrednosti izraza na desni, saj se vrednosti odštejeta, ker sta v vektorju vrednosti prisotni z istima predznakoma. Povezave med vozlišči iz različnih skupin povečujejo vrednost izraza, saj imata vrednosti v vektorju drugačna predznaka.

Predznaki komponent Fiedlerjevega vektorja razdelijo povezan graf na dve skupini. Več kot ima graf povezav med obema skupinama, večja je algebraična povezanost. Bližje kot je algebraična povezanost 0, bolj je razvidna ločnica med dvema skupinama, ki jih dobimo s Fiedlerjevim vektorjem.

Demmel [1] to prikaže na primeru nihanja. Ko je struna napeta, valovi potujejo vzdolž dolžine z določeno frekvenco. Struno si predstavljamo kot množico vozlišč, ki so povezana v zaporedje. Tak graf imenujemo pot na n točkah. Frekvence, ki ustvarjajo valove, so neposredno povezane z lastnimi vrednostmi in njim pripadajočimi lastnimi vektorji Laplaceove matrike. La-

stni vektor tako opisuje gibanje posameznih vozlišč, ki povezana skupaj tvorijo pot. Povezavo med lastnimi vrednostmi in njim pripadajočimi potmi prikazuje slika 2.2.



(a) λ_0. (b) λ_1. (c) λ_2.

Slika 2.2: Slika prikazuje razporeditev vozlišč, ki pripadajo komponentam lastnega vektorja za lastno vrednost λ_i. Najnižja frekvenca je λ_0, λ_1 je druga najnižja frekvenca in λ_2 tretja najnižja.

Druga najmanjša lastna vrednost ustreza stoječemu valu, ki je enak celotni valovni dolžini, kot je prikazano na sliki 2.2(b). To privede do grafa poti s polovico vozlišč pod ravnovesno črto in polovico vozlišč nad njo. Velikost komponente v lastnem vektorju določi višino pripadajočemu vozlišču. Vozliščem, ki so nad sredinsko črto, pripadajo pozitivne komponente Fiedlerjevega vektorja, tistim pod njo pa negativne komponente [14].

Poglavje 3

Metode gručenja

V tem poglavju si bomo ogledali tri metode gručenja. Vse metode bomo opisali, v četrtem poglavju bomo tudi testirali njihovo delovanje, predstavili rezultate in ocenili njihovo uspešnost.

3.1 Metoda voditeljev

Metoda voditeljev je eden najpreprostejših nenadzorovanih učnih algoritmov, ki rešuje problem gručenja. Glavna ideja metode je najti središča posameznih skupin. Z metodo voditeljev je mogoče učinkovito razvrščati tudi do nekaj deset tisoč enot. Postopek metode voditeljev lahko ob različno izbranih začetnih voditeljih pripelje do drugačne rešitve. Zato je potrebno postopek večkrat ponoviti z različnimi začetnimi voditelji [15]. Osnovna shema metode voditeljev je sicer dokaj preprosta in sam postopek implementiranja ni preveč zapleten:

- Izberemo začetno množico voditeljev. Izberemo jih lahko naključno. Izbrati moramo toliko voditeljev, kolikor skupin želimo najti.
- Ponavljamo, dokler se lega voditeljev spreminja:
  - vsak element (v našem primeru komponento Fiedlerjevega vektorja) pripišemo najbližjemu voditelju in tako dobimo skupine,

– za nove voditelje postavimo središča dobljenih skupin.

V našem primeru središča skupin izračunamo kot povprečje komponent v Fiedlerjevem vektorju, ki pripadajo posamezni skupini.

3.1.1 Začetni izbor voditeljev

Začetni izbor voditeljev je možen na več načinov, in sicer [15]:

- **Naključni izbor voditeljev**: Naključni izbor voditeljev lahko pripelje do različnih rešitev, zato je potrebno postopek večkrat ponoviti in na koncu izbrati najboljšo rešitev. Izberemo lahko rešitev, ki se pri veliko ponovitvah pojavi največkrat.
- **Izbor razpršenih voditeljev**: Za voditelja izberemo komponento vektorja, ki je najbolj oddaljena od ostalih. Za drugega voditelja izberemo komponento, ki je najbolj oddaljena od prvega izbranega. Naslednje voditelje izberemo tako, da so čimbolj oddaljeni od že izbranih.

3.1.2 Število voditeljev

Težava nastane pri iskanju števila skupin, ki najbolj ustreza danim podatkom. Pri veliki količini podatkov človek ne more dobro oceniti števila skupin, ki je najbolj primerno za dane podatke. To je ena od slabosti te metode. Na naših podatkih smo čim bolj ustrezno razdelitev grafov na skupine izbrali s pomočjo preizkušanja. Za nekatere grafe smo zato prikazali več možnih razdelitev z različnim številom skupin.

3.1.3 Ustavitveni kriterij

Ker ne moremo z gotovostjo trditi, da se bodo skupine ustalile in nobeno od vozlišč ne bo zamenjalo skupine, moramo postaviti tudi ustavitveni pogoj. Algoritem lahko ustavimo, ko v več korakih zapored manj kot k vozlišč zamenja skupino, ali pa ko se izteče določeno število iteracij.

3.1.4 Metoda voditeljev na naših podatkih

Za potrebe diplomskega dela bomo za dani graf izračunali Fiedlerjev vektor Laplaceove matrike pripadajočega grafa. Na njem bomo izvedli metodo voditeljev. Glede na število iskanih skupin naključno izberemo voditelje. To so kar naključno izbrane komponente Fiedlerjevega vektorja. Vsakemu voditelju pripišemo tiste komponente Fiedlerjevega vektorja, za katere je to najbližji voditelj. Tako dobimo skupine, katerih središča izračunamo kot povprečje ustreznih komponent Fiedlerjevega vektorja. Ta povprečja postanejo novi voditelji in postopek ponavljamo, dokler se skupine ne ustalijo.

3.2 Delitev grafa po predznakih

S pomočjo Fiedlerjevega vektorja lahko razdelimo graf na dve skupini. Vozlišča razporedimo na tista, ki so v Fiedlerjevem vektorju predstavljena z negativnim, ter tista s pozitivnim predznakom. Zato lahko to uporabimo kot metodo gručenja, ki razdeli graf. Za delitev v več kot dve skupini lahko to metodo rekurzivno nadaljujemo na vsaki podskupini. Za vsako podskupino ustvarimo Laplaceovo matriko, ki vsebuje le stolpce in vrstice, ki ustrezajo elementom podskupine. Matrike ne ustvarjamo na novo, ustvarimo le podmatriko matrike celotnega grafa.

Če metodo rekurzivno ponavljamo, metoda teži k številčno čim bolj enakovrednim skupinam, saj smo za ustavitveni kriterij izbrali velikost skupine. Pri deljenju grafa, ki ima dve številčno različni skupini, lahko pride do nepričakovane delitve. Manjša skupina lahko k sebi pritegne kakšno vozlišče, ki bi bolj smiselno pripadalo drugi skupini. Tak primer prikazuje slika 4.11. Do tega pride, ker Fiedlerjev vektor vedno razdeli graf na dve povezani komponenti s polovico vozlišč nad in drugo polovico pod sredinsko črto.

3.2.1 Ustavitveni kriterij

Tudi pri tej metodi moramo izbrati ustavitveni kriterij. Najlažje je, če za ustavitveni pogoj nastavimo največjo dovoljeno velikost skupine. Ko pri deljenju ustvarimo skupino, manjšo od določene velikosti, te skupine ne delimo več. Ostale (večje) skupine rekurzivno delimo naprej. Ko so vse skupine manjše od določene velikosti, se algoritem ustavi.

3.3 Metoda maksimiziranja modularnosti grafa

Modularnost je funkcija $Q(C)$, ki sta jo uvedla Newman and Girvan v delu Finding and evaluating community structure in networks [7]. Modularnost ovrednoti strukturo podgrafa in vsaki skupini c pripiše realno število. Definirana je kot:

$$Q(C) = \frac{1}{2m} \sum_{c \in C} \sum_{i,j \in c} \left(w_{i,j} - \frac{\mathrm{Deg}(i)\,\mathrm{Deg}(j)}{2m} \right),$$

kjer C predstavlja množico skupin c, $w_{i,j}$ predstavlja utež povezave med vozliščema i in j, $\mathrm{Deg}(i)$ pa predstavlja stopnjo vozlišča i. Metoda poskuša vrednost $Q(C)$ maksimizirati in na grafu najti čim bolj gosto povezane skupine [6]. Metoda deluje tako na uteženih kot na neuteženih grafih, števila skupin pa ni potrebno predhodno določiti, saj metoda sama poišče optimalno število skupin. Te metode nismo implementirali sami, smo pa uporabili v Wolfram Mathematici vgrajeno funkcijo *FindGraphCommunities*, ki skupine išče z metodo maksimiziranja modularnosti. Primerjali smo jo z metodo voditeljev ter metodo deljenja po predznakih.

3.4 Mera podobnosti

Za mero podobnosti smo izbrali Randov indeks [12]. Randov indeks je pogosto uporabljena mera za ocenjevanje podobnosti skupin in ocenjevanje podobnosti rezultatov različnih metod razvrščanja. Randov indeks je število

med 0 in 1. Z Randovim indeksom lahko primerjamo dve gručenji podatkov. Bližje kot je indeks 1, bolj se skupine ujemajo. Če poznamo pravilno delitev podatkov, lahko tudi ocenimo uspešnost posamezne metode.

Randov indeks izračunamo po naslednjem postopku. Naj bo dana množica $S = \{s_1, ..., s_n\}$ in dve možni delitvi $U = \{U_1, ..., U_r\}$ ter $V = \{V_1, ..., V_s\}$, kjer imata delitev U in delitev V vsaka po r oziroma s skupin. Definiramo še:

- a = število parov elementov iz S, kjer sta oba elementa v isti skupini v množici U in v isti skupini v množici V,
- b = število parov elementov iz S, kjer sta oba elementa v različni skupini v množici U in v različni skupini v množici V,
- c = število parov elementov iz S, kjer sta oba elementa v isti skupini v množici U in v različni skupini v množici V,
- d = število parov elementov iz S, kjer sta oba elementa v različni skupini v množici U in v isti skupini v množici V.

Randov indeks izračunamo kot:

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}.$$

Primer: Naj bo $S = \{s_1, s_2, s_3, s_4\}$, delitvi pa $U = \{\{s_1\}, \{s_2, s_3, s_4\}\}$ ter $V = \{\{s_1, s_2\}, \{s_3, s_4\}\}$. Tabela 3.1 prikazuje vse možne pare elementov iz S. Dobimo $a = 1$, $b = 2$, $c = 2$ in $d = 1$ ter $\binom{4}{2} = 6$, zato je Randov indeks enak

$$R = \frac{1+2}{6} = \frac{3}{6} = 0,5.$$

Opazimo, da se števec pri računanju Randovega indeksa poveča natanko takrat, ko sta v delitvah U in V oba elementa iz para v isti skupini (poveča se a), ali ko sta v obeh delitvah v različni skupini (poveča se b). Randov indeks se torej poveča natanko takrat, ko se delitvi o nekem paru strinjata.

vsi možni pari	skupini glede na delitev U	skupini glede na delitev V	za 1 povečamo
$\{s_1, s_2\}$	U_1, U_2	V_1, V_1	d
$\{s_1, s_3\}$	U_1, U_2	V_1, V_2	b
$\{s_1, s_4\}$	U_1, U_2	V_1, V_2	b
$\{s_2, s_3\}$	U_2, U_2	V_1, V_2	c
$\{s_2, s_4\}$	U_2, U_2	V_1, V_2	c
$\{s_3, s_4\}$	U_2, U_2	V_2, V_2	a

Tabela 3.1: Tabela prikazuje postopek za izračun Randovega indeksa z vsemi možnimi pari iz $S = \{s_1, s_2, s_3, s_4\}$ in delitvama $U = \{\{s_1\}, \{s_2, s_3, s_4\}\}$ ter $V = \{\{s_1, s_2\}, \{s_3, s_4\}\}$. V drugem in tretjem stolpcu smo z $U = \{\{U_1, U_2\}$, kjer je $U_1 = \{s_1\}$, $U_2 = \{s_2, s_3, s_4\}$ in $V = \{\{V_1, V_2\}$, ker je $V_1 = \{s_1, s_2\}$, $V_2 = \{s_3, s_4\}$ označili, kateri skupini pripada posamezen element iz para.

Poglavje 4

Implementacija

4.1 Programski jezik in okolje

Programerski del diplomskega dela smo napisali v programskem jeziku Python. Ta programski jezik je zelo primeren za reševanje takih nalog, saj lahko uporabljamo različne knjižnice za iskanje lastnih vrednosti matrike. Poleg tega obstajajo tudi že implementirani učni algoritmi, kot je metoda voditeljev. Za razvojno okolje smo izbrali PyCharm. Grafi v diplomskem delu so narisani z Wolfram Mathematico.

4.2 Izbrane metode gručenja

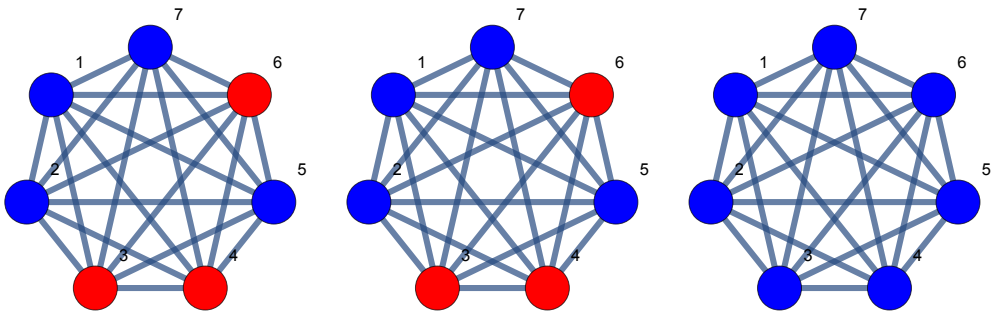
Za metodo voditeljev smo se odločili, ker je učinkovita in dokaj preprosta metoda za gručenje. Uporabili smo jo za gručenje komponent Fiedlerjevega vektorja ter komponentam na koncu pripisali pripadajoča vozlišča. Metodo delitve grafa po predznakih smo izbrali zato, ker je dokaj preprosta in pogosto uporabljena metoda za iskanje skupin na grafih. Za primerjavo smo uporabili še metodo maksimiziranja modularnosti grafa, ki je implementirana v Wolfram Mathematici.

4.3 Testiranje

Za potrebe testiranja smo generirali nekaj posebnih primerov grafov, na katerih smo testirali vse tri metode, opisane v tretjem poglavju. Za nekatere primere smo izračunali tudi Randov indeks in tako ocenili podobnost rešitev z različnimi metodami gručenja.

4.3.1 Polni graf K_n

Polni graf je graf, v katerem je vsako vozlišče povezano z vsemi ostalimi. Polni graf z n vozlišči označimo s K_n. Slika 4.1 prikazuje rezultate vseh treh metod na polnem grafu K_7. Metoda voditeljev in metoda deljenja po predznakih vrneta enak rezultat, medtem ko metoda modularnosti zaradi simetrije grafa in goste povezanosti odkrije le eno skupino. Lastna vrednost, ki pripada Fiedlerjevemu vektorju, je enaka 7.



(a) Metoda voditeljev. (b) Delitev po predznakih. (c) Metoda modularnosti.

Slika 4.1: Slika prikazuje rezultate vseh treh metod na grafu K_7. Randov indeks ob primerjavi metode voditeljev in metode deljenja po predznakih je enak 1, saj sta delitvi enaki. Pri primerjanju metode deljenja po predznakih in metode modularnosti je indeks enak 0,429.

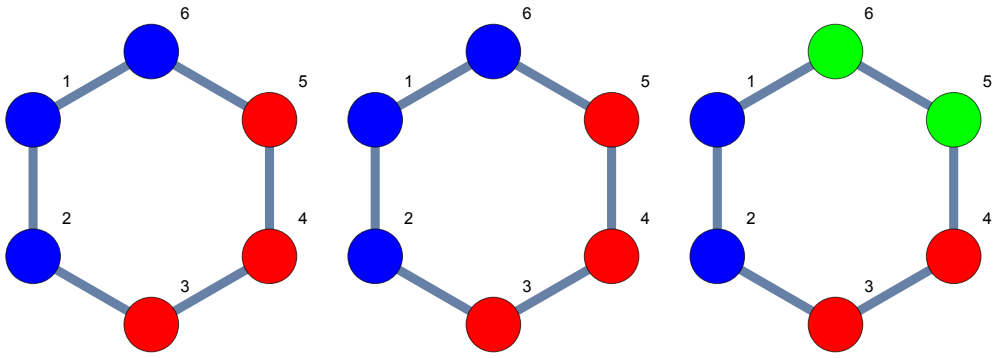
4.3.2 Ciklični graf C_n

Ciklični graf C_n je povezan graf, v katerem je število vozlišč enako številu povezav, vsako vozlišče pa ima stopnjo 2.

Fiedlerjev vektor pri $n = 6$ je enak

$$\begin{bmatrix} -0,53433132 \\ -0,07777841 \\ 0,45655291 \\ 0,53433132 \\ 0,07777841 \\ -0,45655291 \end{bmatrix}.$$

Pri iskanju dveh skupin metoda voditeljev in metoda delitve po predznakih prideta do iste rešitve, metoda modularnosti pa najde na grafu C_6 tri skupine.



(a) Metoda voditeljev. (b) Delitev po predznakih. (c) Metoda modularnosti.

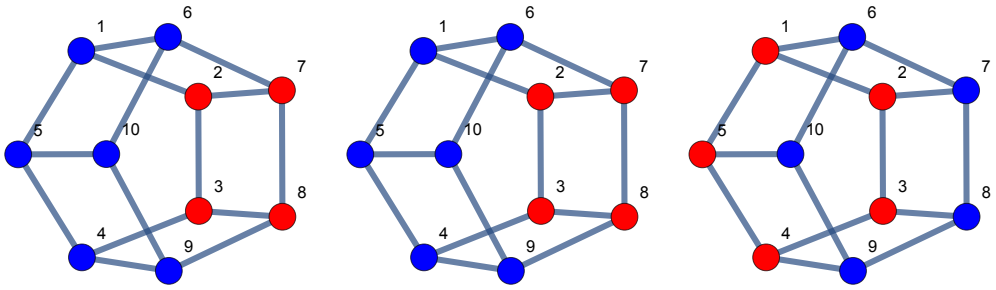
Slika 4.2: Slika prikazuje rezultate vseh treh metod na grafu C_6. Randov indeks pri primerjanju metode voditeljev in metode deljenja po predznakih je enak 1. Ob primerjavi metode deljenja po predznakih in metode modularnosti je Randov indeks enak 0,667.

4.3.3 Prizma Pr_n

Prizma je oglato geometrijsko telo (polieder), omejeno z dvema osnovnima ploskvama in plaščem. Osnovni ploskvi prizme sta skladna in vzporedna

n-kotnika. Plašč je sestavljen iz n paralelogramov, napetih med stranice obeh osnovnih ploskev. Te paralelograme imenujemo stranske ploskve prizme. Stranice obeh osnovnih ploskev imenujemo osnovni robovi prizme. Vse ostale robove imenujemo stranski robovi. Za vse $n \geq 3$ lahko n-strani prizmi priredimo graf Pr_n, ki ima za vozlišča oglišča, za povezave pa osnovne in stranske robove prizme.

Pri iskanju dveh skupin v Pr_5 nas rezultat nekoliko preseneti, saj bi morda človek za vozlišča vsake od skupin izbral kar vozlišča, ki pripadajo posameznemu petkotniku. Vendar pri metodi voditeljev in metodi delitve po predznakih dobimo drugačno rešitev. Prikazana je na sliki 4.3. Rezultat z metodo modularnosti je nekoliko bolj pričakovan.



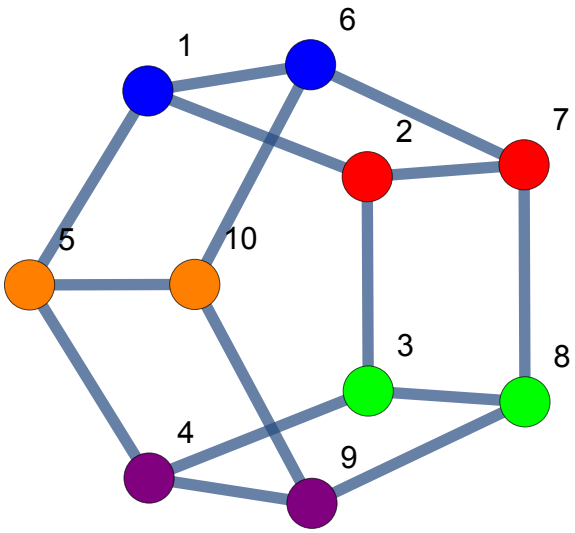
(a) Metoda voditeljev. (b) Delitev po predznakih. (c) Metoda modularnosti.

Slika 4.3: Slika prikazuje rezultate vseh treh metod na Pr_5. Randov indeks pri primerjavi metode deljenja po predznakih in metode modularnosti je 0,444.

Pri iskanju petih skupin metoda voditeljev vrne rešitev, ki se zdi na prvi pogled bolj logična kot pri iskanju dveh skupin. To rešitev prikazuje slika 4.4.

4.3.4 Graf drevo $T(m, n)$

Zvezda je graf z enim središčnim vozliščem (stičiščem) in nekaj listi. Zvezdo z n listi označimo z Z_n. Graf, ki ga dobimo, če povežemo stičišči zvezd Z_m in Z_n, imenujemo drevo $T(m, n)$.



Slika 4.4: Slika prikazuje razdelitev petstrane prizme z metodo voditeljev na pet skupin. Randov indeks pri primerjavi metode voditeljev pri iskanju petih skupin in metode modularnosti je 0,444.

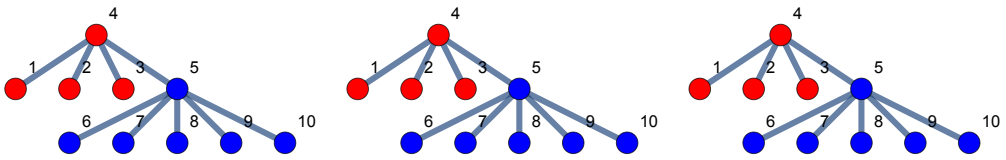
Pričakujemo, da bosta zvezdi tvorili vsaka eno od skupin. Za graf $T(3,5)$ na sliki 4.5 vse tri metode vrnejo enak rezultat. Če graf utežimo, metoda voditeljev pripelje do drugačne rešitve. Takšno rešitev dobimo, ker je Fiedlerjev vektor Laplaceove matrike enak

$$\begin{bmatrix} -0,44851388 \\ -0,44851388 \\ -0,44851388 \\ -0,0993726 \\ 0,06131018 \\ 0,27672081 \\ 0,27672081 \\ 0,27672081 \\ 0,27672081 \\ 0,27672081 \end{bmatrix}.$$

Opazimo, da je komponenta Fiedlerjevega vektorja, ki pripada petemu voz-

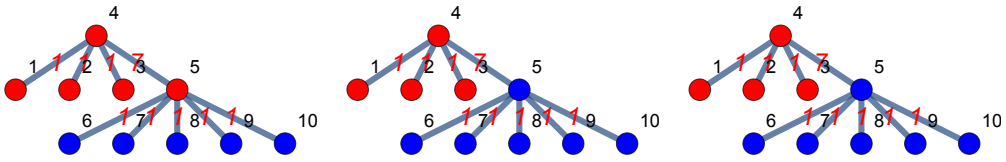
lišču, po velikosti bližje pozitivnim kot negativnim komponentam vektorja, zato ga metoda voditeljev dodeli drugi skupini kot preostali dve metodi. Tudi pri grafu $T(2, 8)$ vse tri metode vrnejo enako rešitev, in sicer vse metode odkrijejo obe zvezdi, kot je prikazano na sliki 4.7.

Tudi pri nekoliko drugačnem grafu, ko med seboj v polni graf povežemo središča štirih zvezd Z_3, vse tri metode vrnejo enako rešitev. Ob dobro nastavljenem parametru števila skupin oziroma največji dovoljeni velikosti skupin vse tri metode za graf $T(3, 3, 3, 3)$ vrnejo rešitev, ki jo prikazuje slika 4.8.



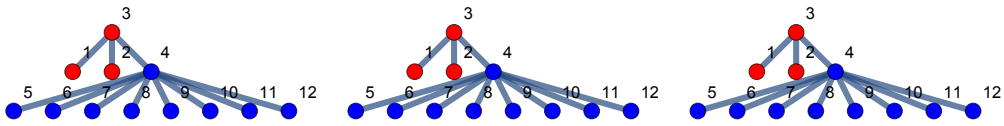
(a) Metoda voditeljev. (b) Delitev po predznakih. (c) Metoda modularnosti.

Slika 4.5: Slika prikazuje rezultate vseh treh metod na grafu $T(3, 5)$. Randov indeks pri primerjavi katerih koli dveh metod je enak 1.



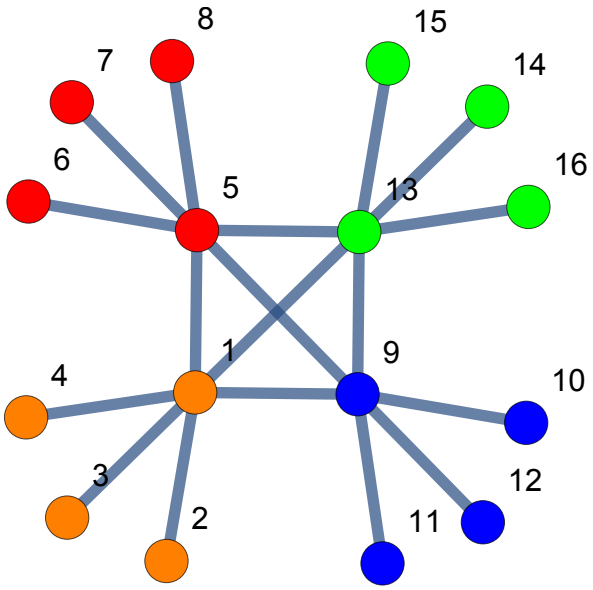
(a) Metoda voditeljev. (b) Delitev po predznakih. (c) Metoda modularnosti.

Slika 4.6: Slika prikazuje rezultate vseh treh metod na uteženem grafu $T(3, 5)$. Na grafu so z rdečimi številkami napisane uteži, s črnimi pa oznake vozlišč. Vse uteži so ena, le utež povezave med stičiščema zvezd Z_3 in Z_5 je 7. Randov indeks primerjave metode voditeljev in metode delitve po predznakih je enak 0,778.



(a) Metoda voditeljev. (b) Delitev po predznakih. (c) Metoda modularnosti.

Slika 4.7: Slika prikazuje rezultate vseh treh metod na neuteženem grafu $T(2, 8)$.



Slika 4.8: Slika prikazuje razdelitev grafa $T(3, 3, 3, 3)$. Tako rešitev vrnejo vse tri metode.

4.3.5 Graf dveh povezanih ciklov $C(m, n)$

Če v vsakem od ciklov C_m in C_n izberemo po eno vozlišče in uniji obeh ciklov dodamo še povezavo med njima, dobimo graf, ki ga označimo s $C(m, n)$. Pričakujemo, da bo na takem grafu vsak od ciklov tvoril eno od skupin. Izkaže se, da različne metode iskanja skupin včasih pripeljejo do različnih rešitev. Za graf $C(4, 4)$ pričakovano vse metode pripeljejo do enake rešitve z dvema skupinama s štirimi vozlišči (slika 4.9). Bolj zanimivi so primeri, kjer

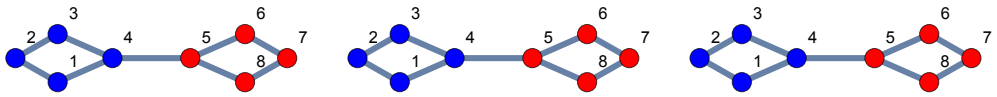
je en cikel manjši od drugega. Na grafu $C(4,7)$ vse metode še vedno pridejo do enakih rešitev in v vsako skupino uvrstijo svoj cikel. Rešitev prikazuje slika 4.10. Drugače je pri grafu $C(4,8)$, kjer metode pridejo do različnih rešitev, ki jih prikazuje slika 4.11. Fiedlerjev vektor grafa $C(4,8)$ je enak

$$\begin{bmatrix} -0,400 \\ -0,438 \\ -0,400 \\ -0,294 \\ -0,032 \\ 0,102 \\ 0,219 \\ 0,298 \\ 0,326 \\ 0,298 \\ 0,219 \\ 0,102 \end{bmatrix}.$$

Ker je peto vozlišče v vektorju po velikosti bliže ciklu z osmimi vozlišči, ga metoda voditeljev uvrsti v skupino s ciklom C_8. Kljub temu ima to vozlišče negativen predznak, zato metoda delitve po predznakih to vozlišče pripiše manjšemu ciklu. Zato sta končna rezultata različna, kar lahko vidimo na slikah 4.11(a) in 4.11(b). Tudi pri grafih z drugačnimi kombinacijami ciklov smo opazili, da do različnih rešitev pride, ko eden od ciklov vsebuje vsaj štiri vozlišča manj od drugega. Primeri takih grafov so $C(1,6), C(2,8), C(30,36)$...

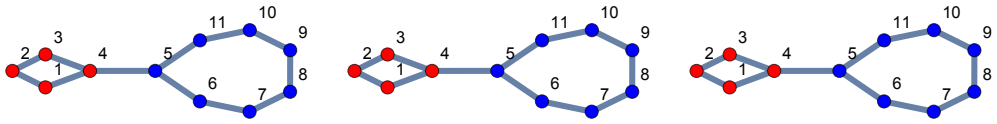
4.3.6 Petersenov graf

Petersenov graf je kubični graf z desetimi vozlišči in petnajstimi povezavami. Vsa vozlišča grafa imajo stopnjo tri. Ima mnogo zanimivih značilnosti in se velikokrat pojavi kot primer in protiprimer pri problemih v teoriji grafov. Imenuje se po danskem matematiku Juliusu Petersenu [5].



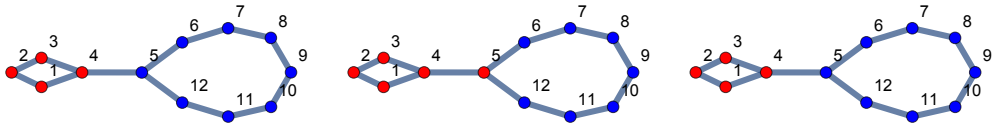
(a) Metoda voditeljev. (b) Delitev po predznakih. (c) Metoda modularnosti.

Slika 4.9: Slika prikazuje rezultate vseh treh metod na grafu $C(4, 4)$. Ker sta cikla enaka, vse tri metode pripeljejo do enakega rezultata.



(a) Metoda voditeljev. (b) Delitev po predznakih. (c) Metoda modularnosti.

Slika 4.10: Slika prikazuje rezultate vseh treh metod na grafu $C(4, 7)$. Vse tri metode pripeljejo do enakega rezultata.



(a) Metoda voditeljev. (b) Delitev po predznakih. (c) Metoda modularnosti.

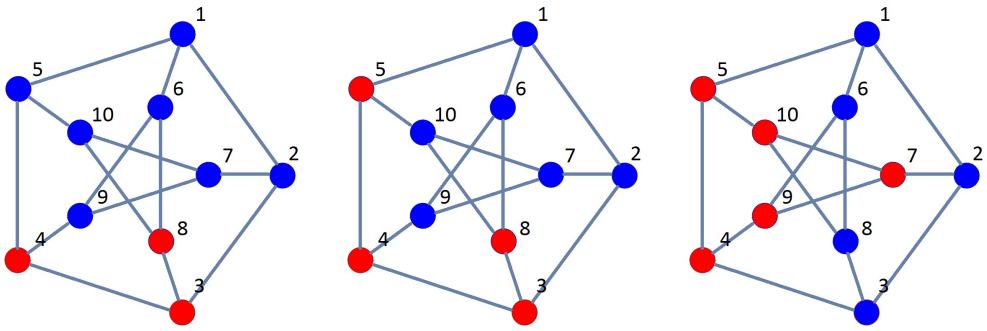
Slika 4.11: Slika prikazuje rezultate vseh treh metod na grafu $C(4, 8)$. Metoda deljenja po predznakih pripelje do drugačnega rezultata, saj ima Fiedlerjev vektor negativen predznak pri petih vozliščih. Metoda voditeljev peto vozlišče razporedi v drugo skupino kljub negativnemu predznaku. Randov indeks primerjave teh dveh gručenj je enak 0,833.

Tudi pri Petersenovem grafu se pojavi podobna situacija kot na primeru

grafa $C(4, 8)$, saj je Fiedlerjev vektor enak

$$\begin{bmatrix} -0.23610671 \\ -0,18520762 \\ 0,55940193 \\ 0,40442929 \\ 0,02442548 \\ -0,07532458 \\ -0,50850283 \\ 0,34018025 \\ -0,17939812 \\ -0,1438971 \end{bmatrix}.$$

Opazimo, da imajo le štiri komponente Fiedlerjevega vektorja pozitiven predznak, in sicer so to komponente, ki pripadajo vozliščem 3, 4, 5 in 8. Vendar pa ima komponenta, ki pripada petemu vozlišču in je sicer negativna, zelo majhno absolutno vrednost. Zaradi tega je kljub negativnemu predznaku dejansko bližje pozitivnim komponentam. Metoda voditeljev to zazna in jo zato dodeli skupini z vozlišči 3, 4, 5 in 8, medtem ko se metoda delitve po predznakh na take izjeme ne ozira. Metoda maksimiziranja modularnosti vrne rešitev z dvema skupinama s petimi vozlišči.



(a) Metoda voditeljev. (b) Delitev po predznakih. (c) Metoda modularnosti.

Slika 4.12: Slika prikazuje rezultate vseh treh metod na Petersenovem grafu. Randova indeksa primerjave rezultatov metode delitve po predznakih s preostalima dvema sta 0,8 (ob primerjavi z metodo voditeljev) in 0,444 (ob primerjavi z metodo maksimiziranja modularnosti). To pomeni, da sta si rezultata metode voditeljev in metode deljenja po predznakih veliko bolj podobna kot rezultata metode deljenja po predznakih in metode maksimiziranja modularnosti.

Poglavje 5

Realni podatki

5.1 Pridobivanje podatkov

V prejšnem poglavju smo algoritme testirali na testnih grafih. Namen je bil spoznati in razumeti delovanje uporabljenih metod. V tem poglavju je opisan postopek pridobivanja podatkov ter rezultat algoritmov na realnih podatkih. Zbrali smo podatke, ki jih najdemo na portalu SICRIS [13]. Osredotočili smo se na raziskovalce Fakultete za računalništvo in informatiko Univerze v Ljubljani (FRI). Zaradi velike količine podatkov smo izbrali obdobje petih let. Iz SICRIS-ove baze smo prenesli podatke za obdobje med letoma 2014 in 2018. Za posameznega raziskovalca smo zbrali le podatke iz dveh sklopov, in sicer:

- 1.01 Izvirni znanstveni članek,
- 1.08 Objavljeni znanstveni prispevek na konferenci.

Tako dobljene podatke smo zapisali v datoteko. Za vsak članek smo najprej poiskali vse avtorje, nato pa izločili tiste, ki niso raziskovalci na FRI. Za vsak članek smo nato poiskali vse možne pare avtorjev tega članka. Nato smo podatke zapisali v uteženo Laplaceovo matriko. Če sta raziskovalca z indeksoma i in j skupaj napisala $w_{i,j}$ člankov, sta vrednosti $l_{i,j}$ in $l_{j,i}$ v Laplaceovi matriki enaki $-w_{i,j}$. Iz te matrike lahko napišemo Laplaceovo

matriko, ki pripada neuteženemu grafu. To naredimo tako, da vse neničelne izvendiagonalne elemente spremenimo v -1, na diagonalo pa namesto vsote uteži zapišemo stopnje vozlišč.

Ker naš algoritem deluje le na povezanih grafih, moramo iz matrike izločiti raziskovalce, ki ne pripadajo največji komponenti grafa. To naredimo s pomočjo knjižnice *networkx*. Iz Laplaceove matrike vseh podatkov generiramo še Laplaceovo matriko največje komponente grafa. Laplaceovo matriko največje komponente bomo uporabili za iskanje skupin.

5.2 Rezultati

Ko ustvarimo Laplaceovo matriko za utežen oziroma neutežen graf, izračunamo Fiedlerjev vektor matrike in začnemo z izvajanjem metode voditeljev ali metode deljenja grafa po predznakih. En primer metode deljenja grafa po predznakih bomo predstavili bolj podrobno, za ostale pa bomo zgolj napisali rezultate ter ocenili smiselnost ustvarjenih skupin.

Podrobno bomo predstavili primer iskanja skupin uteženega grafa z metodo deljenja po predznakih glede na pozitiven ali negativen del lastnega vektorja. Pri tem skupino delimo naprej, če je v posamezni skupini več kot 15 raziskovalcev, v nasprotnem primeru pa skupino označimo kot končno. To število smo določili eksperimentalno po več poskusih delitev in primerjanju Randovega indeksa rezultatov.

Začetna skupina je sestavljena iz 87 raziskovalcev. Pri prvem deljenju dobimo dve še vedno veliki skupini, eno s 30 in drugo s 57 raziskovalci.

V drugem koraku obe skupini razpadeta na manjše podskupine. Dobimo štiri skupine, ki jih prikazuje tabela 5.1. V zgornji vrstici je navedena zaporedna številka skupine, v oklepaju pa velikost skupine. Skupina 1.1 vsebuje osem raziskovalcev in je zato že skupina, ki je ne delimo naprej. Za preostale tri skupine ustvarimo nove Laplaceove matrike tako, da za vsako skupino ustvarimo podmatriko matrike celotnega grafa. Ta podmatrika je sestavljena iz vrstic in stolpcev, ki ustrezajo raziskovalcem v posamezni skupini. Nato

ponovno izvedemo deljenje vsake od skupin na dva dela.

Skupina 1.2 v tretjem koraku razpade na skupini 1.2.1 in 1.2.2. Skupina 2.1 razpade na 2.1.1, 2.1.2, skupina 2.2 pa na 2.2.1 in 2.2.2. Vseh šest novo ustvarjenih skupin v tretjem koraku delitve prikazuje tabela 5.2.

Vse razen dveh skupin vsebujejo manj kot petnajst raziskovalcev, zato v četrtem koraku razdelimo le skupini 1.2.2 in 2.1.1. Deljenje teh dveh skupin prikazuje tabela 5.3. Končne skupine so prikazane na sliki 5.1 in v tabeli 5.4.

Ob podrobnejšem pregledu rezultatov opazimo, da smo z deljenjem grafov dobili ne samo skupine raziskovalcev, ki so med seboj sodelovali, temveč deloma tudi laboratorije na fakulteti. To je bilo pričakovano, saj raziskovalci znotraj laboratorijev sodelujejo bolj pogosto. Zato smo na sliko pred zaporedno številko vsakega raziskovalca zapisali številko, ki predstavlja laboratorij na FRI, katerega član je ta raziskovalec.

Številke, dodeljene laboratorijem, so:

- 1: Laboratorij za adaptivne sisteme in paralelno procesiranje,
- 2: Laboratorij za algoritmiko,
- 3: Laboratorij za bioinformatiko,
- 4: Laboratorij za biomedicinske računalniške sisteme in oslikave,
- 5: Laboratorij za e-medije,
- 6: Laboratorij za informatiko,
- 7: Laboratorij za integracijo informacijskih sistemov,
- 8: Laboratorij za kognitivno modeliranje,
- 9: Laboratorij za kriptografijo in računalniško varnost,
- 10: Laboratorij za matematične metode v računalništvu in informatiki,
- 11: Laboratorij za podatkovne tehnologije,

- 12: Laboratorij za računalniške komunikacije,
- 13: Laboratorij za računalniške strukture in sisteme,
- 14: Laboratorij za računalniški vid,
- 15: Laboratorij za računalniško grafiko in multimedije,
- 16: Laboratorij za tehnologijo programske opreme,
- 17: Laboratorij za umetne vizualne spoznavne sisteme,
- 18: Laboratorij za umetno inteligenco,
- 19: Laboratorij za vseprisotne sisteme.

Če uporabimo neutežen graf z največjo dovoljeno velikostjo skupine 15, dobimo delitev, ki jo prikazuje slika 5.2. Opazimo, da večina članov v posamezni skupini pripada enemu ali dvema laboratorijema. Nekatere skupine imajo tudi raziskovalce iz več laboratorijev. Za vsako skupino s slike 5.2 smo zapisali, katerim laboratorijem pripadajo raziskovalci uvrščeni v to skupino. Prva številka predstavlja laboratorij, številka v oklepaju pa število raziskovalcev iz tega laboratorija, uvrščenih v skupino:

- RUMENA SKUPINA: 18(5), 3(2) in 2(1),
- ZELENA SKUPINA: 6(2) in 16(2),
- ORANŽNA SKUPINA: 14(6), 8(5) in 7(1),
- TEMNO VIJOLIČNA SKUPINA: 2(5), 16(2), 1(1) in 14(1),
- RDEČA SKUPINA: 8(6), 1(1), 2(1), 6(1), 12(1), 13(1) in 18(1),
- MODRA SKUPINA: 13(4), 3(1) in 17(1),
- SINJA SKUPINA: 11(8),
- ROZA SKUPINA: 3(7), 1(3) in 8(2),

- ČRNA SKUPINA: 17(8) in 10(1),
- SVETLO VIJOLIČNA SKUPINA: 15(5), 3(1) in 11(1).

Pri uporabi metode voditeljev so bili rezultati nekoliko drugačni. Uporabili smo metodo voditeljev pri iskanju devetih skupin. Dobljeno razporeditev na neuteženem grafu kaže slika 5.4, na uteženem pa slika 5.3. V tabelah 5.4, 5.6 in 5.8 so zapisane vse končne skupine na uteženih grafih. Končne skupine na neuteženih grafih pa prikazujejo tabele 5.5, 5.7 in 5.9. Vsaka tabela prikazuje rezultate z eno od metod.

5.3 Primerjava rezultatov na obstoječih rešitvah

Oglejmo si, kako se metoda voditeljev in metoda deljenja grafa po predznakih primerjata z metodo maksimiziranja modularnosti grafa. Rešitev, ki jo dobimo z uporabo funkcije *FindGraphCommunities* v Wolfram Mathematici, smo prikazali na slikah 5.5 in 5.6.

5.4 Ocena uspešnosti programa

Za oceno uspešnosti programa smo uporabili Randov indeks. Primerjali smo najdene skupine z laboratoriji, ki jim raziskovalci pripadajo:

- gručenje na sliki 5.1 (utežen graf, deljenje po predznakih): 0,908,
- gručenje na sliki 5.2 (neutežen graf, deljenje po predznakih): 0,907,
- gručenje na sliki 5.3 (utežen graf, metoda voditeljev): 0,894,
- gručenje na sliki 5.4 (neutežen graf, metoda voditeljev): 0,917,
- gručenje na sliki 5.5 (utežen graf, metoda maksimiziranja modularnosti): 0,859,

- gručenje na sliki 5.6 (neuteženen graf, metoda maksimiziranja modularnosti): 0,922.

Na neuteženem grafu smo izračunali Randov indeks tudi za vse kombinacije vseh treh metod. Randov indeks ob primerjavi gručenja z metodo deljenja grafa po predznakih in metodo voditeljev je 0,931. Randov indeks ob primerjavi gručenj metode deljenja po predznakih in metode maksimiziranja modularnosti je 0,959. Ob primerjavi gručenj metode voditeljev in metode maksimiziranja modularnosti je Randov indeks enak 0,943. Vrednosti Randovega indeksa so blizu ena, razlike med njimi so majhne, to pomeni, da so vse tri metode pripeljale do podobne razdelitve skupin.

Metoda deljenja po predzankih in metoda voditeljev sta na uteženem in neuteženem grafu vrnili podobne rezultate. Skupine, ki jih vrne metoda modularnosti na uteženem grafu pa se precej razlikujejo od skupin na neuteženem grafu.

Ob primerjavi gručenj na uteženih in neutženih grafih opazimo, da je metoda deljenja po predznakih na obeh grafih vrnila zelo podobne skupine. Na uteženem grafu dobimo devet, na neuteženem pa deset skupin. Kljub temu so pari posameznih vozlišč večinoma v istih skupinah. Randov indeks ob primerjavi rezultatov metode na uteženem in neuteženem grafu je 0,968. Zato sta tudi Randova indeksa primerjav teh dveh gručenj z laboratoriji na FRI podobna.

Tudi pri metodi voditeljev so skupine na uteženem in neuteženem grafu primerljive. Randov indeks ob primerjavi gručenj na uteženem in neuteženem grafu je enak 0,941. Tudi Randova indeksa gručenj uteženega in neuteženega grafa ob primerjavi z laboratoriji na FRI sta si podobna. Randov indeks je enak 0,894 ob primerjavi laboratorijev z gručenjem na uteženem in 0,917 ob primerjavi z gručenjem na neuteženem grafu.

Drugače je pri metodi maksimiziranja modularnosti. Tu opazimo razliko v številu skupin, saj metoda na uteženem grafu vrne deset, na neuteženem pa le osem skupin. Randov indeks ob primerjavi gručenj na teh dveh grafih je enak 0,907. Randov indeks primerjave laboratorijev z gručenjem z metodo

modularnosti na uteženem grafu je enak 0,859. Ob primerjavi laboratorijev in gručenja na neuteženem grafu pa je Randov indeks enak 0,922. Opazimo, da je razlika precej večja kot pri ostalih dveh metodah, saj se najdene skupine na uteženem in neuteženem grafu med seboj precej razlikujejo.

1.1 (8)	1.2 (22)	2.1 (34)	2.2 (23)
2.28365	3.37827	1.16109	1.19515
3.16324	3.50664	1.31379	2.04646
3.35424	10.19284	1.33385	2.12766
18.02275	11.51331	1.33795	2.18188
18.20389	13.05957	2.14989	2.22475
18.29020	13.13442	3.12536	2.23400
18.33187	13.29198	3.23399	6.21393
18.35423	13.52244	3.30142	6.31407
	15.15677	3.30921	7.50004
	15.16131	3.35422	8.27674
	15.30062	3.37515	8.36491
	15.35071	3.38461	8.36914
	15.52309	6.15294	14.09581
	17.05896	8.04242	14.11161
	17.18198	8.14565	14.19226
	17.29381	8.15295	14.22472
	17.30155	8.28779	14.23401
	17.34398	8.29485	14.31252
	17.39227	8.29486	14.32887
	17.50002	8.31563	16.03307
	17.50367	8.34600	16.13564
	17.50843	8.38174	16.23953
		8.39226	16.25527
		11.16154	
		11.20334	
		11.25526	
		11.30918	
		11.33188	
		11.34156	
		11.35425	
		11.36470	
		12.52331	
		13.21404	
		18.37516	

Tabela 5.1: Štiri skupine po dveh deljenjih grafa po predznakih.

1.2.1 (5)	1.2.2 (17)	2.1.1 (19)	2.1.2 (15)	2.2.1 (10)	2.2.2 (13)
3.37827	3.50664	1.16109	1.33795	7.50004	1.19515
13.05957	10.19284	1.31379	2.14989	8.27674	2.04646
13.13442	11.51331	1.33385	6.15294	8.36491	2.12766
13.29198	15.15677	3.12536	8.04242	8.36914	2.18188
13.52244	15.16131	3.23399	8.14565	14.09581	2.22475
	15.30062	3.30142	8.15295	14.11161	2.23400
	15.35071	3.30921	8.28779	14.19226	6.21393
	15.52309	3.35422	8.29485	14.22472	6.31407
	17.05896	3.37515	8.29486	14.31252	14.23401
	17.18198	3.38461	8.31563	14.32887	16.03307
	17.29381	8.38174	8.34600		16.13564
	17.30155	11.16154	8.39226		16.23953
	17.34398	11.20334	12.52331		16.25527
	17.39227	11.25526	13.21404		
	17.50002	11.30918	18.37516		
	17.50367	11.33188			
	17.50843	11.34156			
		11.35425			
		11.36470			

Tabela 5.2: Tabela prikazuje delitev skupin 1.2, 2.1 in 2.2, vsaka od njih razpade na dve manjši skupini. Naprej bomo delili le skupini 1.2.2 in 2.1.1, ostale skupine pa imajo po 15 ali manj raziskovalcev.

1.2.2.1 (10)	1.2.2.2 (7)	2.1.1.1 (11)	2.1.1.2 (8)
10.19284	3.50664	1.16109	11.16154
17.05896	11.51331	1.31379	11.20334
17.18198	15.15677	1.33385	11.25526
17.29381	15.16131	3.12536	11.30918
17.30155	15.30062	3.23399	11.33188
17.34398	15.35071	3.30142	11.34156
17.39227	15.52309	3.30921	11.35425
17.50002		3.35422	11.36470
17.50367		3.37515	
17.50843		3.38461	
		8.38174	

Tabela 5.3: Skupina 1.2.2 razpade na 1.2.2.1, 1.2.2.2, skupina 2.1.1 pa na 2.1.1.1 in 2.1.1.2.

1.1 (8)	1.2.1 (5)	2.1.2 (15)	2.2.1 (10)	2.2.2 (13)	1.2.2.1 (10)	1.2.2.2 (7)	2.1.1.1 (11)	2.1.1.2 (8)
2.28365	3.37827	1.33795	7.50004	1.19515	10.19284	3.50664	1.16109	11.16154
3.16324	13.05957	2.14989	8.27674	2.04646	17.05896	11.51331	1.31379	11.20334
3.35424	13.13442	6.15294	8.36491	2.12766	17.18198	15.15677	1.33385	11.25526
18.02275	13.29198	8.04242	8.36914	2.18188	17.29381	15.16131	3.12536	11.30918
18.20389	13.52244	8.14565	14.09581	2.22475	17.30155	15.30062	3.23399	11.33188
18.29020		8.15295	14.11161	2.23400	17.34398	15.35071	3.30142	11.34156
18.33187		8.28779	14.19226	6.21393	17.39227	15.52309	3.30921	11.35425
18.35423		8.29485	14.22472	6.31407	17.50002		3.35422	11.36470
		8.29486	14.31252	14.23401	17.50367		3.37515	
		8.31563	14.32887	16.03307	17.50843		3.38461	
		8.34600		16.13564			8.38174	
		8.39226		16.23953				
		12.52331		16.25527				
		13.21404						
		18.37516						

Tabela 5.4: Tabela vseh končnih skupin z metodo deljenja po predznakih na uteženem grafu, največja dovoljena velikost skupine je 15.

1 (8)	2 (4)	3 (12)	4 (9)	5 (12)	6 (6)	7 (8)	8 (12)	9 (9)	10 (7)
2.28365	6.21393	7.50004	1.19515	1.33795	3.37827	11.16154	1.16109	10.19284	3.50664
3.16324	6.31407	8.15295	2.04646	2.14989	13.05957	11.20334	1.31379	17.18198	11.51331
3.35424	16.03307	8.27674	2.12766	6.15294	13.13442	11.25526	1.33385	17.29381	15.15677
18.02275	16.25527	8.34600	2.18188	8.04242	13.29198	11.30918	3.12536	17.30155	15.16131
18.20389		8.36491	2.22475	8.14565	13.52244	11.33188	3.23399	17.34398	15.30062
18.29020		8.36914	2.23400	8.29485	17.05896	11.34156	3.30142	17.39227	15.35071
18.33187		14.09581	14.23401	8.29486		11.35425	3.30921	17.50002	15.52309
18.35423		14.11161	16.13564	8.31563		11.36470	3.35422	17.50367	
		14.19226	16.23953	8.39226			3.37515	17.50843	
		14.22472		12.52331			3.38461		
		14.31252		13.21404			8.28779		
		14.32887		18.37516			8.38174		

Tabela 5.5: Tabela vseh končnih skupin z metodo deljenja po predznakih na neuteženem grafu, največja dovoljena velikost skupine je 15.

1 (4)	2 (7)	3 (19)	4 (13)	5 (5)	6 (10)	7 (10)	8 (11)	9 (8)
13.05957	3.50664	1.19515	7.50004	6.15294	10.19284	8.28779	1.16109	2.28365
13.13442	11.51331	1.33795	8.15295	8.04242	17.05896	8.38174	1.31379	3.16324
13.29198	15.15677	2.04646	8.27674	8.14565	17.18198	11.16154	1.33385	3.35424
13.52244	15.16131	2.12766	8.34600	8.31563	17.29381	11.20334	3.12536	18.02275
	15.30062	2.14989	8.36491	12.52331	17.30155	11.25526	3.23399	18.20389
	15.35071	2.18188	8.36914		17.34398	11.30918	3.30142	18.29020
	15.52309	2.22475	14.09581		17.39227	11.33188	3.30921	18.33187
		2.23400	14.11161		17.50002	11.34156	3.35422	18.35423
		6.21393	14.19226		17.50367	11.35425	3.37515	
		6.31407	14.22472		17.50843	11.36470	3.37827	
		8.29485	14.23401				3.38461	
		8.29486	14.31252					
		8.39226	14.32887					
		13.21404						
		16.03307						
		16.13564						
		16.23953						
		16.25527						
		18.37516						

Tabela 5.6: Tabela vseh končnih skupin z metodo voditeljev na uteženem grafu pri iskanju devetih skupin.

1 (10)	2 (10)	3 (7)	4 (13)	5 (10)	6 (11)	7 (10)	8 (7)	9 (9)
10.19284	2.28365	3.50664	8.28779	1.16109	2.04646	7.50004	8.15295	1.19515
17.05896	13.05957	11.51331	8.38174	1.31379	2.12766	8.27674	8.29485	1.33795
17.18198	13.13442	15.15677	11.16154	1.33385	2.22475	8.36491	8.29486	2.14989
17.29381	13.29198	15.16131	11.20334	3.12536	2.23400	8.36914	8.34600	2.18188
17.30155	13.52244	15.30062	11.25526	3.16324	6.21393	14.09581	8.39226	6.15294
17.34398	18.02275	15.35071	11.30918	3.23399	6.31407	14.11161	13.21404	8.04242
17.39227	18.20389	15.52309	11.33188	3.30142	16.03307	14.19226	18.37516	8.14565
17.50002	18.29020		11.34156	3.30921	16.13564	14.22472		8.31563
17.50367	18.33187		11.35425	3.35422	16.23953	14.23401		12.52331
17.50843	18.35423		11.36470	3.35424	16.25527	14.31252		
				3.37515		14.32887		
				3.37827				
				3.38461				

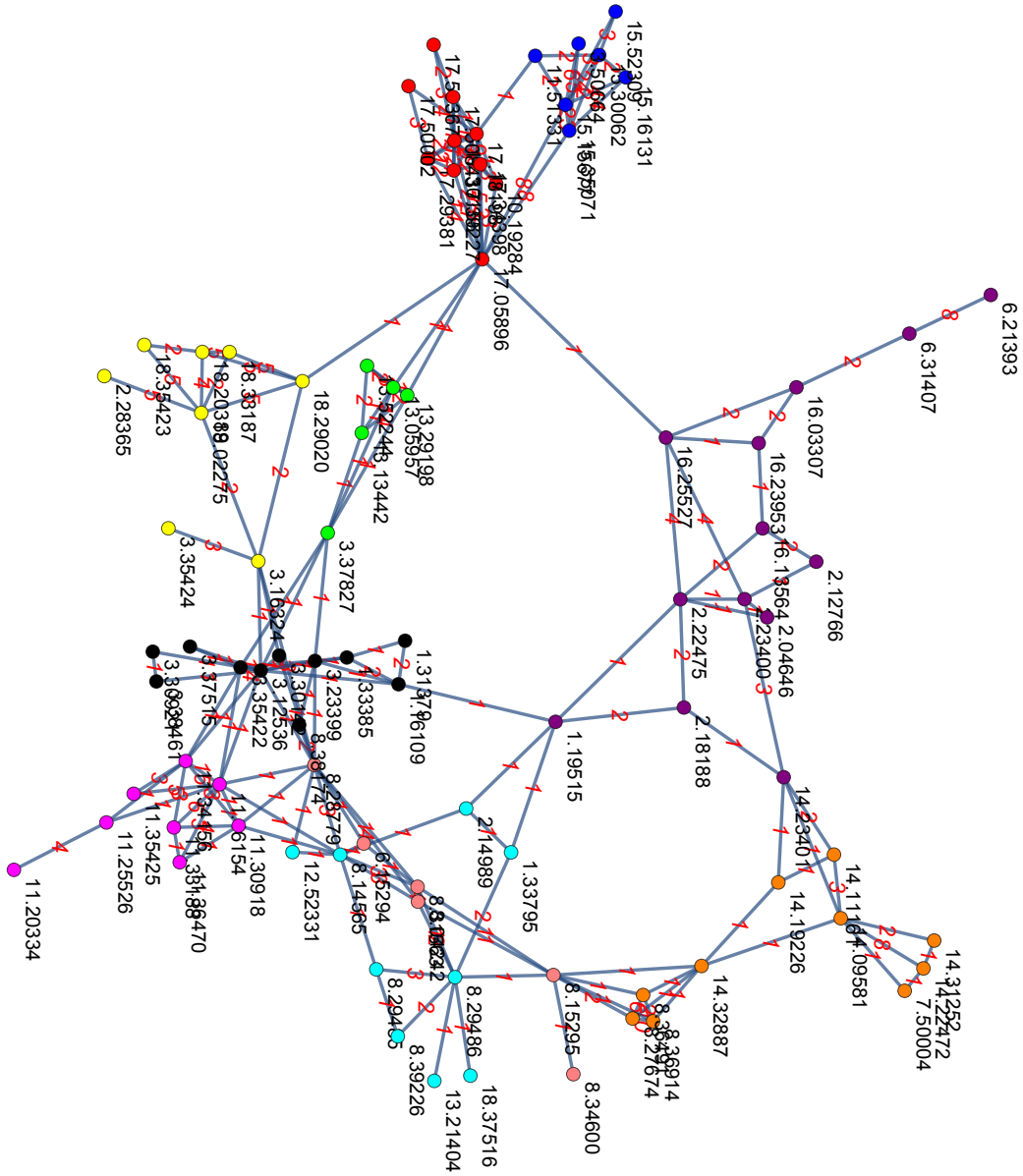
Tabela 5.7: Tabela vseh končnih skupin z metodo voditeljev na neuteženem grafu pri iskanju devetih skupin.

1 (14)	2 (12)	3 (11)	4 (10)	5 (8)	6 (8)	7 (7)	8(7)	9 (6)	10 (4)
1.33795	1.19515	1.16109	10.19284	11.16154	2.28365	7.50004	3.50664	8.15295	13.05957
2.14989	2.04646	1.31379	17.05896	11.20334	3.16324	14.09581	11.51331	8.27674	13.13442
6.15294	2.12766	1.33385	17.18198	11.25526	3.35424	14.11161	15.15677	8.34600	13.29198
8.04242	2.18188	3.12536	17.29381	11.30918	18.02275	14.19226	15.16131	8.36491	13.52244
8.14565	2.22475	3.23399	17.30155	11.33188	18.20389	14.22472	15.30062	8.36914	
8.28779	2.23400	3.30142	17.34398	11.34156	18.29020	14.23401	15.35071	14.32887	
8.29485	6.21393	3.30921	17.39227	11.35425	18.33187	14.31252	15.52309		
8.29486	6.31407	3.35422	17.50002	11.36470	18.35423				
8.31563	16.03307	3.37515	17.50367						
8.38174	16.13564	3.37827	17.50843						
8.39226	16.23953	3.38461							
12.52331	16.25527								
13.21404									
18.37516									

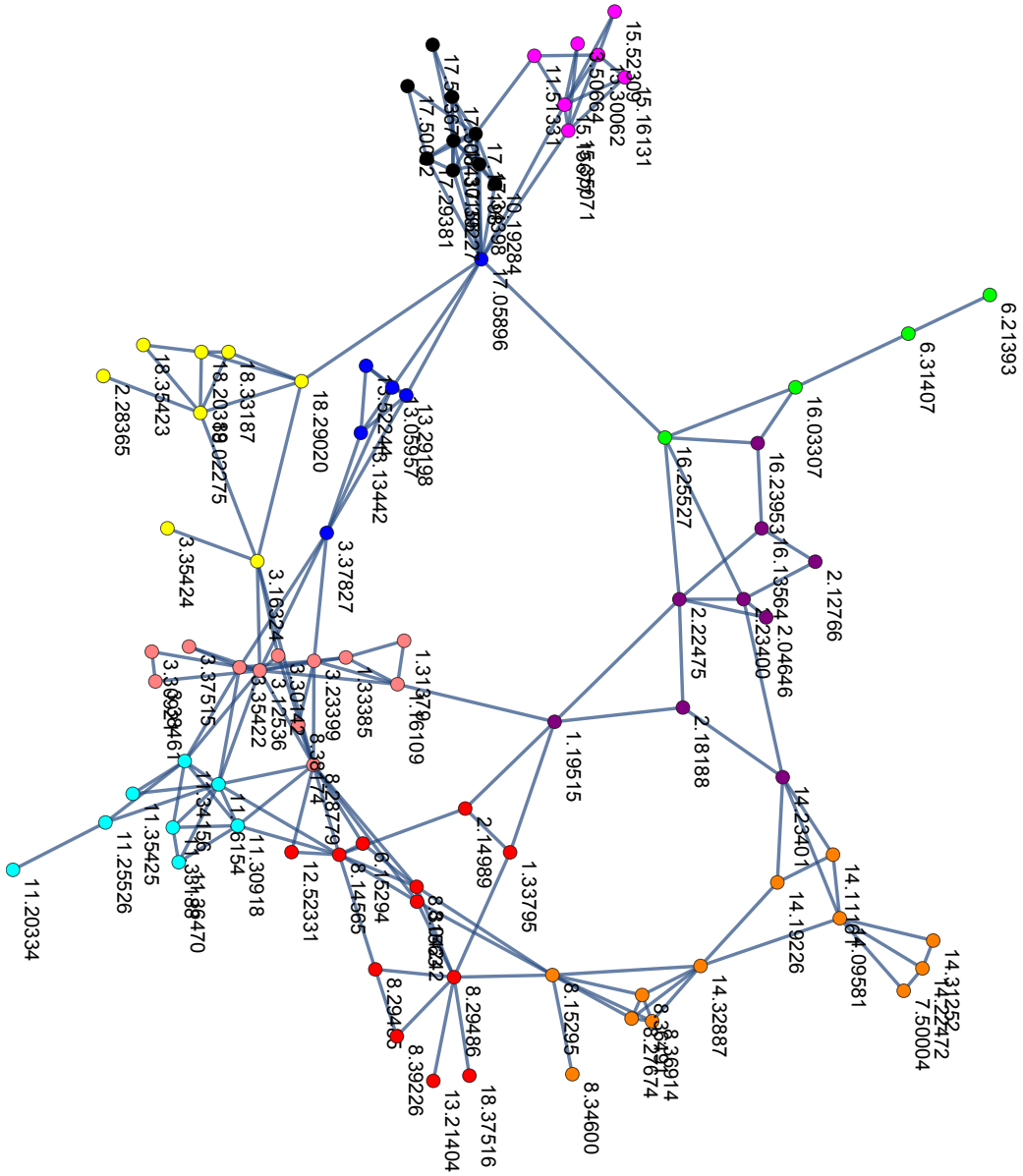
Tabela 5.8: Tabela vseh končnih skupin z metodo maksimiziranja modularnosti na uteženem grafu.

1 (26)		2 (14)	3 (13)	4 (10)	5 (8)	6 (7)	7 (5)	8(4)
1.16109	8.14565	1.19515	7.50004	10.19284	2.28365	3.50664	8.29485	13.05957
1.31379	8.28779	1.33795	8.15295	17.05896	3.16324	11.51331	8.29486	13.13442
1.33385	8.31563	2.04646	8.27674	17.18198	3.35424	15.15677	8.39226	13.29198
3.12536	8.38174	2.12766	8.34600	17.29381	18.02275	15.16131	13.21404	13.52244
3.23399	11.16154	2.14989	8.36491	17.30155	18.20389	15.30062	18.37516	
3.30142	11.20334	2.18188	8.36914	17.34398	18.29020	15.35071		
3.30921	11.25526	2.22475	14.09581	17.39227	18.33187	15.52309		
3.35422	11.30918	2.23400	14.11161	17.50002	18.35423			
3.37515	11.33188	6.21393	14.19226	17.50367				
3.37827	11.34156	6.31407	14.22472	17.50843				
3.38461	11.35425	16.03307	14.23401					
6.15294	11.36470	16.13564	14.31252					
8.04242	12.52331	16.23953	14.32887					
		16.25527						

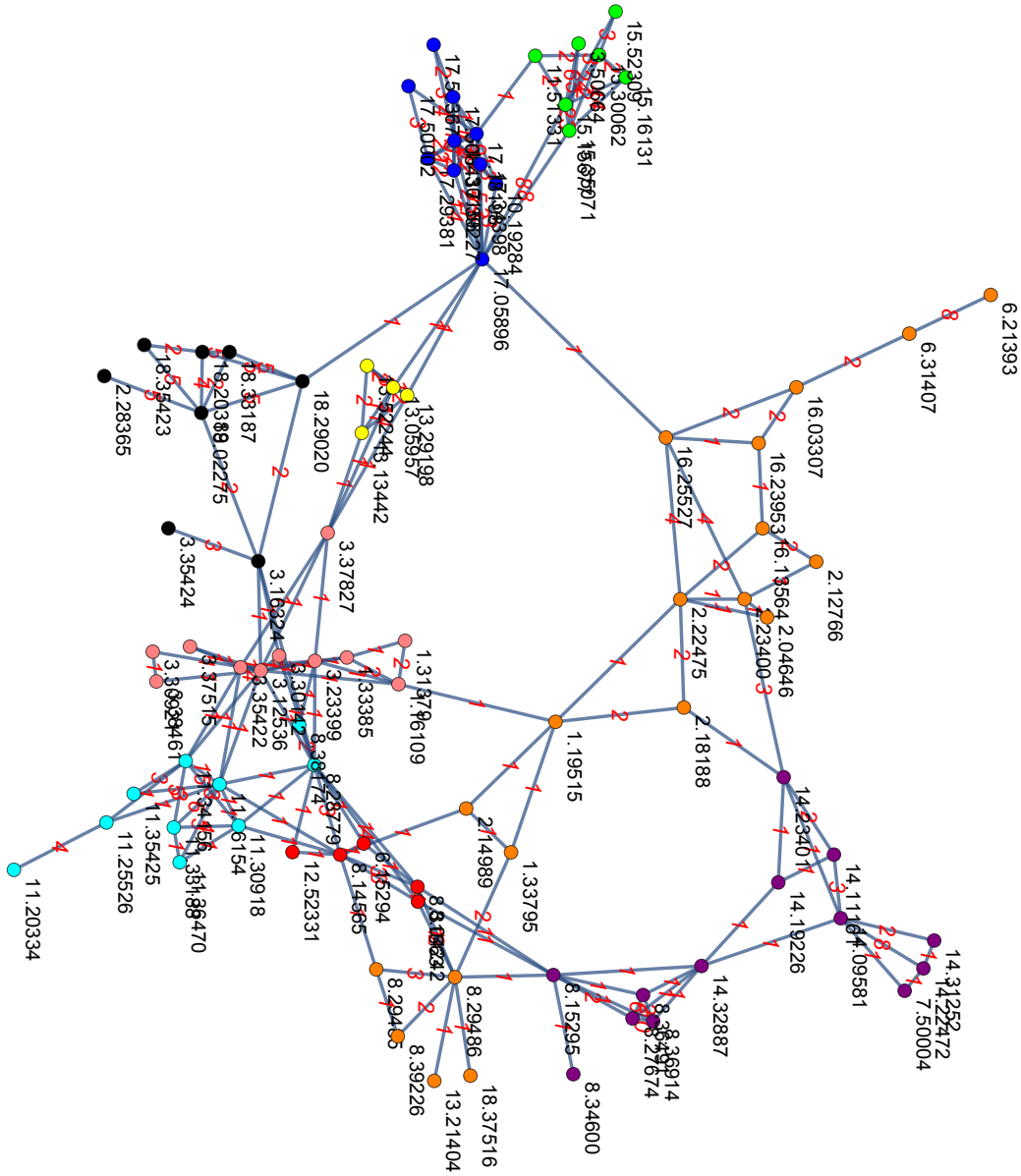
Tabela 5.9: Tabela vseh končnih skupin z metodo maksimiziranja modularnosti na neuteženem grafu.



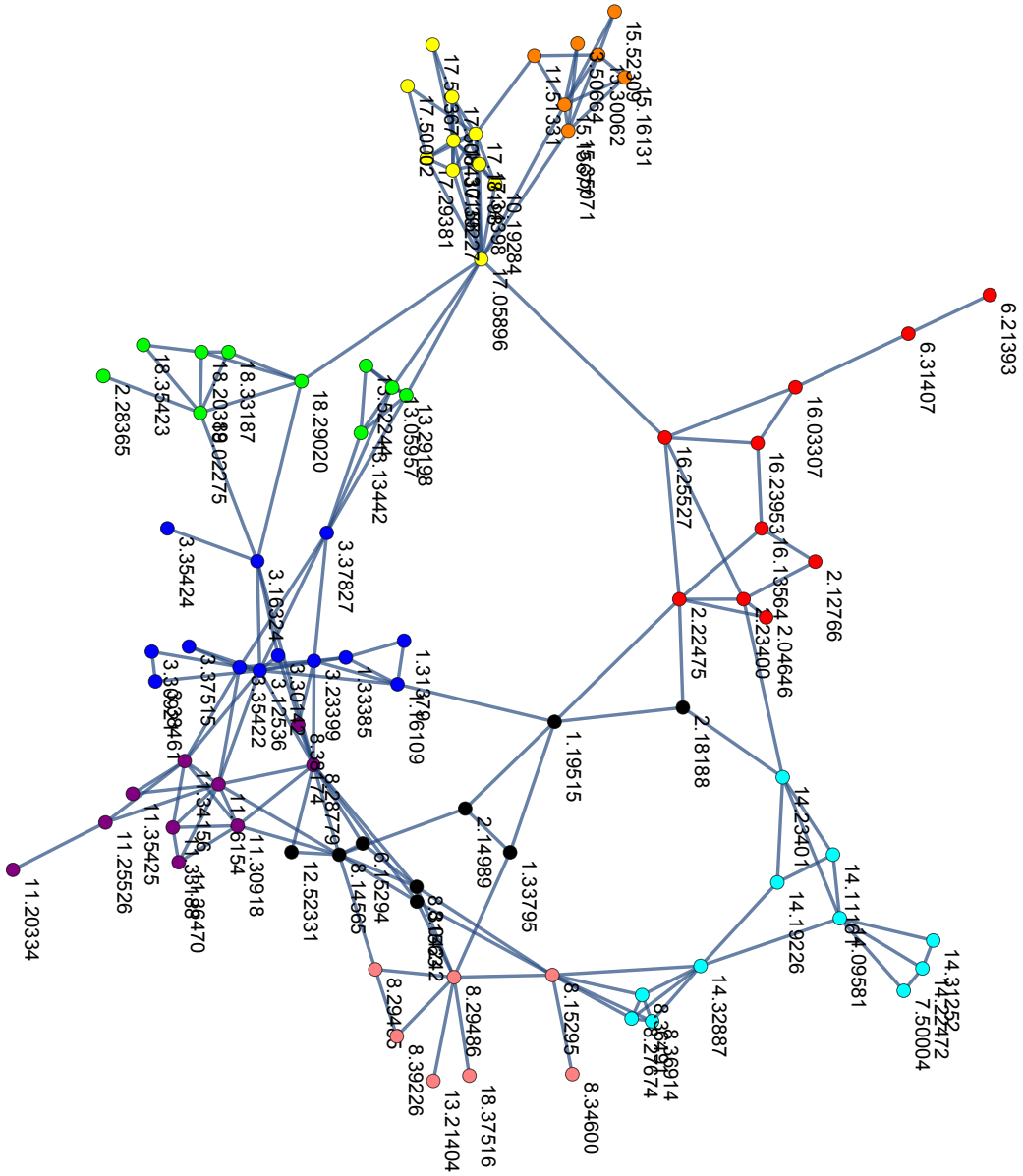
Slika 5.1: Slika prikazuje deljenje uteženega grafa z metodo deljenja po predznakih, največja dovoljena velikost skupine je 15. Končne skupine prikazane na grafu so zapisane tudi v tabeli 5.4.



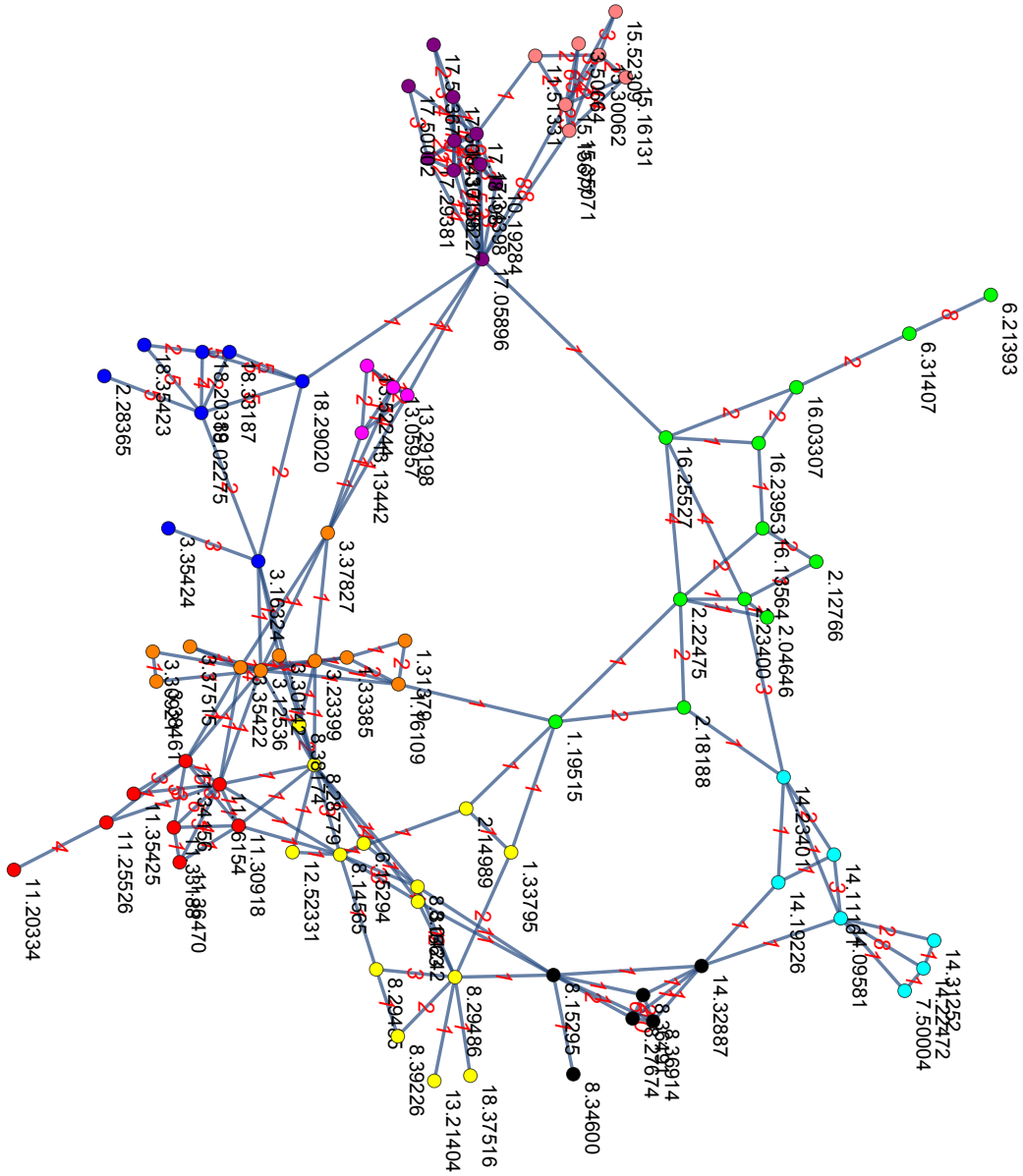
Slika 5.2: Slika prikazuje deljenje neuteženega grafa z metodo deljenja po predznakih, največja dovoljena velikost skupine je 15. Končne skupine prikazane na grafu so zapisane tudi v tabeli 5.5.



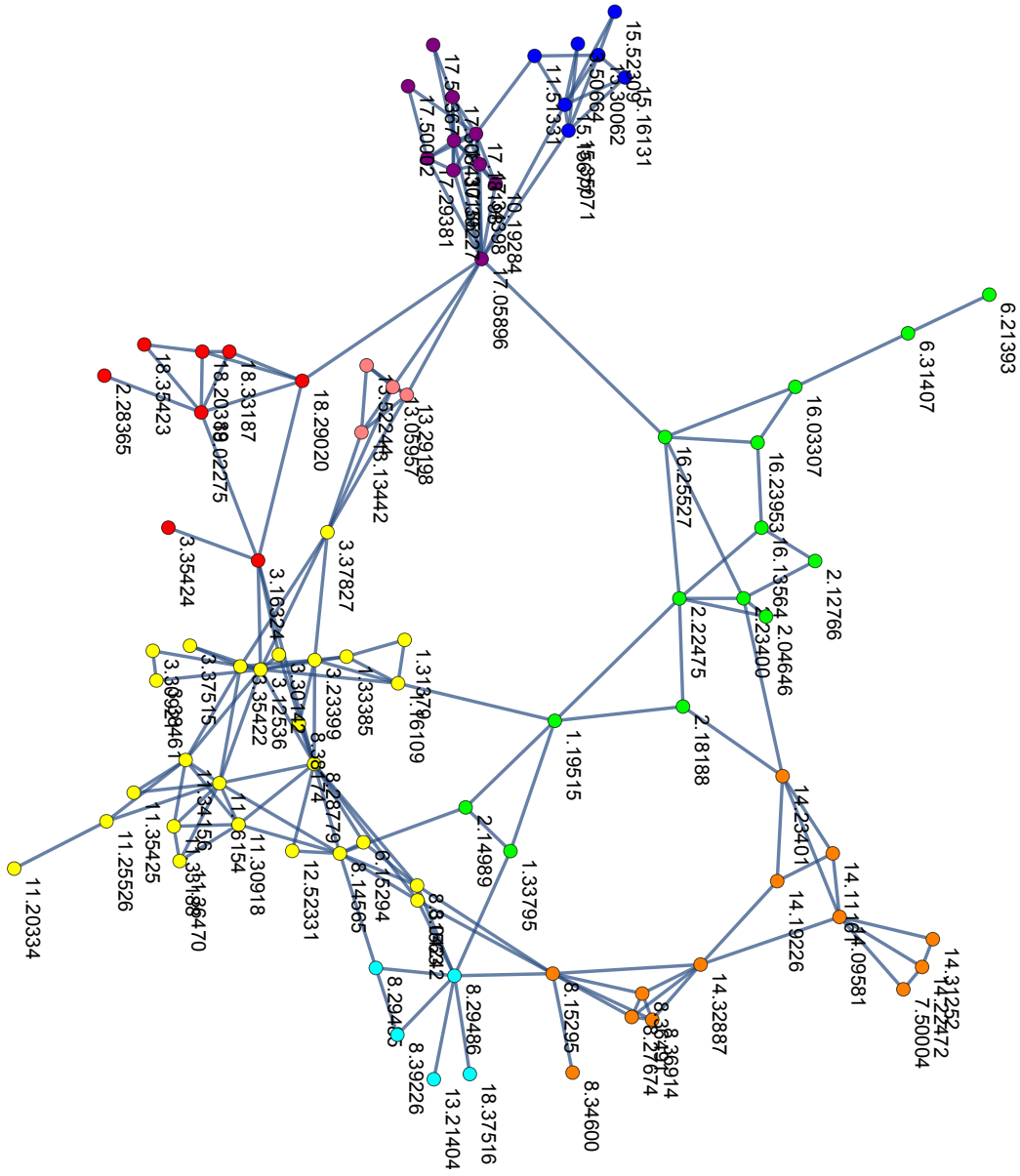
Slika 5.3: Slika prikazuje deljenje uteženega grafa z metodo voditeljev pri iskanju devetih skupin. Končne skupine prikazane na grafu so zapisane tudi v tabeli 5.6.



Slika 5.4: Slika prikazuje deljenje neuteženega grafa z metodo voditeljev pri iskanju devetih skupin. Končne skupine prikazane na grafu so zapisane tudi v tabeli 5.7.



Slika 5.5: Najdene skupine na uteženem grafu s funkcijo *FindGraphCommunities* v Wolfram Mathematici. Končne skupine prikazane na grafu so zapisane tudi v tabeli 5.8.



Slika 5.6: Najdene skupine na neuteženem grafu s funkcijo *FindGraphCommunities* v Wolfram Mathematici. Končne skupine prikazane na grafu so zapisane tudi v tabeli 5.9.

Poglavje 6

Zaključek

Cilj diplomskega dela je bilo preučevanje Laplaceove matrike ter njene druge najmanjše lastne vrednosti in pripadajočega lastnega vektorja. Opisali smo lastni vektor, ki pripada najmanjši lastni vrednosti Laplaceove matrike ter Fiedlerjev vektor. Opisali smo tri metode gručenja, to so metoda voditeljev, metoda delitve grafa po predznakih ter metoda maksimiziranja modularnosti.

Napisali smo program, ki najde skupine v grafu in temelji na deljenju komponent Fiedlerjevega vektorja z metodo voditeljev ali metodo deljenja po predznakih. Za primerjavo smo uporabili metodo maksimiziranja modularnosti v funkciji *FindGraphCommunities* v Wolfram Mathematici. Na nekaj majhnih primerih smo prikazali delovanje našega programa in vseh treh metod.

Na koncu smo delovanje preverili na realnih podatkih, kjer smo skupine iskali med raziskovalci na FRI. Ustvarili smo graf, ki temelji na podatkih o člankih iz SICRIS-ove baze podatkov. Program je iskal skupine tako na uteženem kot na neuteženem grafu. Rezultate smo primerjali s pomočjo Randovega indeksa. Opazili smo, da najdene skupine deloma sovpadajo z laboratoriji na FRI. Program smo primerjali z obstoječo rešitvijo *FindGraphCommunities* ter ugotovili, da dobimo primerljive rezultate.

Literatura

[1] James Demmel. Graph partitioning, part 2. Dosegljivo: `https://people.eecs.berkeley.edu/~demmel/cs267/lecture20/lecture20.html`, 1999. [Dostopano: 20.11.2018].

[2] Gašper Fijavž. *Diskretne strukture*. Založba FRI, 2017.

[3] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486:75–174, 2010.

[4] Radu Horaud. A short tutorial on graph laplacians, laplacian embedding, and spectral clustering. Dosegljivo: `https://csustan.csustan.edu/~tom/Clustering/GraphLaplacian-tutorial.pdf`, 2017. [Dostopano: 3.3.2019].

[5] Robin J. Wilson in John J. Watkins. *Uvod v teorijo grafov*. Društvo matematikov, fizikov in astronomov Slovenije, 1997.

[6] Maria Cristina Vasconcelos Nascimento in Leonidas S. Pitsoulis. Community detection by modularity maximization using GRASP with path relinking. *Computers and Operations Research*, 40:3121–3131, 2013.

[7] M. Newman, M. E. J. in Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, 2004.

[8] Mark Newman. *Networks: An Introduction*. Oxford University Press, Inc., 2010.

[9] Bojan Orel. *Linearna algebra*. Založba FRI, 2017.

[10] Shriphani Palakodety. The smallest eigenvalues of a graph Laplacian. Dosegljivo: http://blog.shriphani.com/2015/04/06/the-smallest-eigenvalues-of-a-graph-laplacian/, 2018. [Dostopano: 3.12.2018].

[11] Tatjana Povše. Graf - podatkovna struktura. Dosegljivo: http://www2.nauk.si/materials/417/out-627771/index.html#state=3, 2005. [Dostopano: 7.11.2018].

[12] Jorge M. Santos and Mark J. Embrechts. On the use of the adjusted rand index as a metric for evaluating supervised classification. *Artificial Neural Networks*, 5769:175–184, 2009.

[13] SICRIS. Dosegljivo: http://www.sicris.si, 2019. [Dostopano: 10.1.2019].

[14] Brian Slininger. Fiedler's theory of spectral graph partitioning. Dosegljivo: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.592.1730&rep=rep1&type=pdf, 2013. [Dostopano: 20.11.2018].

[15] Blaž Zupan. Uvod v odkrivanje znanj iz podatkov. Dosegljivo: http://file.biolab.si/textbooks/uozp/zapiski.pdf, 2017. [Dostopano: 18.11.2018].