UNIVERZITET U BEOGRADU
FIZIČKI FAKULTET

MAGISTARSKI RAD

NALAŽENJE OTEŽINJENIH PODSTRUKTURA U NEKIM REALNIM I KOMPJUTERSKI-GENERISANIM KOMPLEKSNIM MREŽAMA

Student: Marija Mitrović

Mentor: Prof. dr Bosiljka Tadić

Beograd, 2010

Zahvalnost za veliki deo svog znanja o teoriji kompleksnih mreža i fizike uopšte dugujem svom mentoru, Prof. dr Bosiljki Tadić. Saradnja sa njom na radovima koje smo zajednički objavile omogućila mi je ne samo da naučim mnogo o fizici kompleksnih sistema, već i o metodologiji naučnog rada, kao i o istraživačkom poštenju koje je neophodno u naučnom radu. Uz svog mentora i komentroa Prof. dr Aleksandra Belića naučila sam veći deo onoga što znam o primenama numeričkih metoda u fizici. Diskusije sa Prof. dr Aleksandrom Bogojevićem omogućile su mi da proširim svoje znanje i vidike u ostalim oblastima fizike i nauke uopšte. Veliku zahvalnost dugujem i svom prijatelju i kolegi dr Antunu Balažu koji je uvek imao vremena da sasluša i odgovori na svako moje pitanje.

Želela bih da se zahvalim i svojim dragim kolegama u Laboratoriji za primenu računara u nauci na stvorenoj fantastičnoj radnoj atmosferi i prijatnom okruženju u laboratoriji koje su neophodne za uspešan istraživački rad.

Naravno veliku zhvalnost dugujem svojoj porodici posebno svojim roditeljima koji su na sve načine podržali i ispratili svaki moj profesionalni i životni korak. Hvala mojim prijateljima koji su mi pružali moralnu podršku kada mi je bila neophodna.

I na kraju zahvaljujem se Laboratoriji za primenu računara u nauci kao i Institutu za fiziku iz Beograda bez čijih resursa i materijalne podrške moj istraživački rad ne bi bio mogući. Zahvaljujem se i Institutu Jožef Stefan i Odseku za teorijsku fiziku iz Ljubljane na pruženom gostoprimstvu tokom mojih boravaka u Sloveniji.

Sadržaj

1 Uvod **9**
1.1 Kompleksne mreže 9
1.1.1 Topološke osobine kompleksnih mreža 10
1.2 Moduli i metodi za njihovo nalaženje 11

2 Metod maksimalne verodostojnosti **15**
2.1 Generalizovani metod 15
2.1.1 Metod maksimalne verodostojnosti za binarne grafove 19
2.2 Numerička implementacija algoritma 20

3 Modularna struktura kompjuterski generisanih grafova **23**
3.1 Binarni modularni grafovi i MMV 23
3.2 Otežinjeni slučajni grafovi 29

4 Modularna struktura realnih otežinjenih mreža **33**
4.1 Mreža genskih ekspresija pivskog kvasca 34
4.2 IMDB 38
4.3 Osobine mreža socijalne zajednice bloga B92 43

5 Zaključak **49**

A Programski kod **51**

Abstrakt

Realni dinamički sistemi mogu se efikasno reprezentovati kompleksnim mrežama čije su strukturne osobine u direktnoj vezi sa dinamikom i evolucijom samih sistema. Kompleksne mreže često ispoljavaju modularnu strukturu, sa podgrafovima različitih veličina i strukture. Razvijanje metoda za nalaženje podgrafova u mrežama i njihova primena su od velikog značaja za bolje razumevanju dinamike i evolucije kompleksnih sistema. U ovoj tezi biće predstavljen generalizovani metod maksimalne verodostojnosti za nalaženje podgrafova u otežinjenim i binarnim grafovima kao i njegova implementacija. Prvo ćemo opisati karakteristike metoda i testirati njegovu efikasnost na kompjuterski generisanim modularnim mrežama. Zatim ćemo pokazati kako se metod može primeniti za nalaženje podgrafova u mrežama generisanim iz realnih podataka. Konkretno, metod će biti primenjen na mreži genksih ekspresija pivskog kvasca kao i na socijalnim mrežama u sajber prostoru što govori o svestranost teorije kompleksnih mreža a samim tim i metoda maksimalne verodostojnosti.

Glava 1

Uvod

Jedna od najizraženijih karakteristika mnogih prirodnih i sistemima čiji je tvorac čovek je njihova kompleksnost. Čak i najjednostavnije forme života, ćelije, zavise od stotina komplikovanih biohemijskih reakcija [1, 2], gde produkt jedne reakcije predstavlja substrat neke druge, koja je opet katalizirana enzimom generisanim u nekoj trećoj reakciji. Niži oblici života sadrže hiljade ćelija koje medjusobno komuniciraju, dok kompleksnost postaje skoro nepojmljiva kada se razmatraju mreže neurona koje sadrže izmedju 10^{15} i 10^{16} veza. Kompleksnost nije karakteristična samo za žive sisteme. Procenjeni broj stranica na vebu je reda veličine 10^{11} sa hiljadama milijardi linkova koji ih povezuju. Drugi primeri velikih kompleksnih sistema kreiranih od strane čoveka su transportne mreže [3] ili internet [4, 5]. I sami ljudi su deo kompleksnih mreža različitih socijalnih odnosa, [6], koji postoje izmedju skoro sedam milijardi individua na Zemlji.
Iako kompleksnost zaokuplja pažnju ljudi već hiljadama godina, potpuno razumevanje kompleksnih sistema ostaje otvoreno pitanje i izazov. Zajedničko za sve gore opisane sisteme je da se sastoje od odredjenog broja jedinica koje medjusobno interaguju često jakim nelinearnim interakcijama. Kao posledica ovih interakcija sistem ispoljava kompleksno kolektivno ponašanje. Postojanje ili odsustvo odredjene interakcije izmedju dva konstituenta sistema je jedna od njegovih fundamentalnih karakteristika. Ova osobina predstavlja osnov za primenu teorije kompleksnih mreža (poznate i kao teorija grafova) kao univerzalnog i sistematskog okvira za izučavanje kompleksnosti karakteristične za veliki broj problema.

1.1 Kompleksne mreže

Kompleksne mreže su sastavljene od čvorova, numerisanih sa i ($i = 1 \dots N$), medjusobno povezanih linkovima e_{ij}. Mreža veličine N se formalno može zapisati u obliku matrice povezanosti A čiji je element $A_{ij} = 1$ ukoliko postoji link od čvora i ka čvoru

j odnosno $A_{ij} = 0$ u suprotnom. Svaka mreža ima jedinstvenu matricu povezanosti do na permutacije indeksa čvorova. Čvorovi u mrežama odgovaraju osnovnim jedinicama gradje realnih sistema kao što su proteini, ćelije, veb stranice, individue, [1], dok se linkovima opisuju interakcije izmedju jedinica, na primer protein-protein interakcije, sinapse koje povezuju neurone, hiperlinkovi, poznanstva. Linkovi mogu biti neusmereni odnosno usmereni i/ili binarni odnosno otežinjeni sve u zavisnosti od tipa interakcije koju predstavljaju. Binarne neusmerene mreže se reprezentuju simetričnim matricama povezanosti, $A_{ij} = A_{ji}$, dok usmrenost linkova indukuje nesimetričnu formu matrice. Za reprezentovanje otežinjenih mreža neophodno je uvesti matricu težina W koja se može posmatrati i kao generalizovana matrica povezanosti. Element matrice $W_{ij} = w$ ($w \neq 0$ i $w \in R$) označava link od čvora i ka čvoru j jačine (težine) w. Ukoliko je $w = 0$ čvor i ne utiče na čvor j. Predstavljanje kompleksnih sistema skupom čvorova i linkova je pojednostavljena slika koja nam pruža mogućnost da niz različitih problem koji se tiču ovih sistema tretiramo na isti način korišćenjem teorije kompleksnih mreža i metodima statističke fizike [1, 7].
Teorija grafova je oblast matematike koja se bavi regularnim, abstraktnim konstrukcijama, koje imaju malo dodira sa realnim mrežama. Značajan napredak u teoriji grafova desio se u pedesetim godinama prošloga veka kada su Erdoš i Renji započeli sa izučavanjem slučajnih grafova [8]. Posebno interesovanje je izazvao model slučajnih grafova, čije su osobine kao sistematski izučavane, [1, 7, 9] i koji je nekoliko decenija predstavljao jedini model realnih mreža. Iako sam model ne poseduje osobine svojstvene mrežama kojima su reprezentovani empirijski podaci, poslužio je kao osnova za definisanje veličina kojima se opisuju topološke osobine mreža, kao što su *raspodela povezanosti*, *klastering koeficijent* ili *srednje rastojanje izmedju čvorova u mreži*. Kompjuterski resursi i dostupnost ogromnim elektronskim bazama krajem devedesetih, uticali su na razvoj oblasti kompleksnih mreža i pojavu prvih modela sa karakteristikama koje odstupaju od karakteristika slučajnih grafova [9, 10, 11].

1.1.1 Topološke osobine kompleksnih mreža

Izučavanju topoloških osobina grafova i njihov veza sa dinamikom i ponašanjem reprezentovanog sistema, posvećena je velika pažnja [7, 1]. Oblik raspodele povezanosti ili srednja vrednost najkraćeg rastojanja su osobine po kojima se kompleksne mreže dobijene iz realnih podataka razlikuju od slučajnih grafova. Ove mreže su karakterisane nehomogenošću njihove strukture na različitim skalama [1]. Lokalna nehomogenost se ogleda u tome da čvorovi u mrežama nisu jednaki u odnosu na broj prvih suseda, q_i. Jedna od veličina kojom se opisuje lokalna stuktura mreže je raspodela povezanosti, $P(q)$, koja predstavlja verovatnoću da slučajno izabrani čvor ima stepen q, odnosno q prvih suseda. Raspodela povezanosti većine realnih kompleksnih mreža ima oblik stepenog zakona, $P(q) \sim q^{-\gamma}$, zbog čega se ove mreže često

nazivaju i *scale free* mrežama [1]. Većina realnih mreža je karakterisana eksponentom γ čije su vrednosti u intervalu $[2, 3]$. Modeli kojima se reprodukuju ova osobina relanih mreža [9, 11] ukazali su na vezu izmedju preferencijalnog povezivanja i *scale free* strukture mreže.
Većina realnih mreže su male, u smislu srednjeg rastojanja izmedju dva čora u mreži [1, 10]. Putanja izmedju i_1 i i_n se definiše kao skup čvorova $\{i_1, i_2, \ldots, i_n\}$ takav da izmedju svaka dva susedna čvora i_k i i_{k+1} postoji link, dok je dužina putanje jednaka broju linkova izmedju ovih čvorova. Jasno je da u mreži izmedju svaka dva čvora postoji više putanja različite dužine. Dužina najkraće putanje koja povezuje dva čvora, l, predstavlja rastojanje izmedju njih, dok je srednje rastojanje, $< l >$, definisano kao srednja vrednost l usrednjena po svakom paru čvorova u mreži. U realnim mrežama srednje rastojanje raste logaritamski sa veličinom mreže, što za posledicu ima da su svi čvorovi u grafu u odnosu jedni na druge na relativno malim rastojanjima [10].
Za većinu socijalnih mreža (na primer prijateljstava) je karakteristično da su prijatelji jedne individue ujedno i medjusobno prijatelji. Ova osobina mreže se opisuje klastering koeficijentom i govori o verovatnoći da su susedi slučajno odabranog čvora takodje povezani. Klastering koeficijent čvora i definisane je formulom $C_i = \frac{e_i}{\frac{1}{2}q_i(q_i-1)}$, dok klastering koeficijent mreže C predstavlja srednju vrednost klasteringa svih njenih čvorova. Model opisan u radu [11] ukazuje na vezu izmedju klasterisanost mreže i procesa prevezivanja linkova.

1.2 Moduli i metodi za njihovo nalaženje

Navedene veličine karakterišu mreže na nivou čvorova (na primer stepen) i na nivou čitave mreže. Medjutim, u većini realnih mreža postoje nehomogenosti i na mezoskopskom nivou. Ove nehomogenosti se manifestuju kao grupe čvorova koji su povezani na specifičan način, poznatih kao podgrafovi. Podgrafovi igraju važnu ulogu kako u strukturi tako i u funkciji same mreže [1, 12], pri čemu različiti tipovi podgrafova karakterišu različite funkcionalne mreže. U zavisnosti od veličine i načina linkovanja čvorova, možemo razlikovati više tipova podgrafova: *motivi* koji se mogu naći u genetskim i komunikacionim mrežama [13], *putanje* i *drveta* koja se pojavljuju kao relevantni podgrafovi u metaboličkim i neuronskim mrežama [14], ili *lanci* koji predstavljaju *motive* tipične za mreže reči u knjigama i mreže električnih vodova [15]. *Zajednice ili moduli* su vrlo važna mezoskopska struktura i usko su povezane sa funkcionalnim karakteristikama samog sistema, na primer grupe ljudi koji interaguju medjusobno u društvu i na vebu [16, 17, 18, 19], veb stranice koje se odnose na istu tematiku [20] ili proteini povezani sa metastazama razlicitih oblika raka [21]. Identifikacija modula odnosno zajednica je od velikog značaja u naporima

da se razume uticaj strukture mreže na njenu dinamiku kao i njena evolucija. Osim toga, može nam pružiti neophodne informacije o pojedinačnim ulogama čvorova u mreži. Na primer, pojedini čvorovi u modulu mogu imati uloge *konektora* preko kojih je modul povezan sa ostatkom mreže ili ulogu *centralnih* čvorova koji kontrolišu i stabilizuju samu zajednicu [22].
Identifikacija modularne strukture mreže predstavlja jedan od najčešće izučavanih problema oblasti teorije grafova što za posledicu ima brz razvoj ove podoblasti [1, 23]. Kada se govori o metodama nalaženja podgrafova u mrežama mora se imati u vidu da ne postoji jedinstvena striktna definicija modula ili zajednice. Od izbora definicije zavisi i izbor metoda kao i širina njegove primene.
Intuitivno, zajednica ili modul u slučaju binarnih grafova se mogu posmatrati kao skup čvorova koji su *gusto* povezani medjusobno a retko sa ostatkom mreže. Formalna definicija zajednice je kompleksan zadatak. Izbor definicije zavisi od odgovora na sledeća pitanja: da li se definicija bazira na lokalnim ili globalnim karakteristikama čvorova, da li dozvoljava da čvor bude deo više zajednica istovremeno, zatim da li se težine linkova uzimaju u obzir, i da li definicija dozvoljava postojanje hijerarhijske strukture zajednica unutar mreže. Globalni metodi uzimaju u obzir strukturu čitave mreže a razlikuju se po karakteristikama koje su značajne za identifikaciju modula. Kao globalni metodi javljaju se algoritmi bazirani na različitim optimizacionim tehnikama, maksimizacija modularnosti [24] ili metod spinskih stakala [25], zatim algoritmi bazirani na različitim centralnostima čvorova i linkova u grafu [16, 26], spektralni i metodi bazirani na sinronizaciji, [27, 28, 29, 30] ili dinamici slučajnih šetnji [31]. Neretko se pojavljuju i metodi koji su kombinacija dva ili više navedenih metoda. Kod lokalinh metoda definicija zajednice zavisi isključivo od lokalnih karakteristika mreže. Primer lokalnih metoda su k-klik perkolacioni metod [32] ili metod koji se bazira na nalaženju prirodnih zajednica oko čvorova optimizacijom fitnes funkcije [33].
U principu čvor može biti deo jednog ili više modula istovremeno, kada se kaže da postoji prepokrivanje modula ili zajednica. Prepokrivanje u zajednicama je najviše izraženo u socijalnim mrežama. Svega par metoda je u stanju da detektuju zajednice ovakvog tipa. Jedan od njih je i metod maksimalne verodostojnosti [34], čija je generalizacija predstavljena u ovoj tezi. U nekim kompleksnim mrežama postoji netrivijalna organizacija zajednica, zajednice zadovoljavaju odredjenu hijerarhijska pravila, što za posledicu ima da ne postoji jedinstvena podela grafa.
Veliki broj algoritama za detekciju *zajednica* je baziran *maksimizaciji modularnosti*, funkciji koja je prvi put uvedena u radu [24]. Funkcija modularnost zavisi od razlike broja linkova koji se nalaze unutar jednog modula i broja linkova koji ih povezuju, odnosno od broja podgrafova kao i njihove strukture. Maksimalna vrednost modularnosti odgovara optimalnoj podeli mreže u G modula. Modularnost, kao i većina funkcija definisanih na kompleksnim sistemima, ima veliki broj lokalnih minimuma. Kao posledicu imamo da vreme potrebno za nalaženje maksimalne vrednosti mod-

ularnosti ne raste polinomijalno sa rastom mreže što problem čini numerički zahtevnim. U cilju da se ovaj algoritam prilagodi i unapredi u poslednjih par godina izvršene su različite modifikacije ovog metoda [28, 35].
Sinhronizacija čvorova u mreži je tesno povezana sa njenom strukturom. Jedan od metoda za nalaženje modula se bazira na faznoj sinhronizaciji Kuramotovih oscilatora [30]. Pokazano je da najpre dolazi do sinhronizacije oscilatora (čvorova) u modulima pa tek onda do sinhronizacije na nivou čitave mreže. U metodu se koristi dinamička matrica povezanosti, $D_t(T)$, čiji je element $D_t(T)$ jednak 1 ukoliko je korelacija izmedju oscilatora i i j veća od T odnosno 0 u svim ostalim trenucima. Za odgovarajuću vrednosti parametra T matrica $D_t(T)$ se sastoji od blokova 1 koji predstavljaju gusto povezane jako korelisane čvorove koji pripadaju istom modulu. Ovaj model dobro radi na relativno gusto povezanim mrežama čiji moduli prema strukturi liče na slučajne grafove, t.j. homogeniji su u smislu lokalnih osobina čvorova. Većina realnih mreža su retke sa vrlo izarženom hijerarhijskom strukturom izmedju čvorova, zbog čega potpuna sinhronizacija čvorova na mreži nije moguća. Dodatni problem metoda predstavlja i nepostojanje kriterijuma za pogodan izbor vrednosti parametra T.
Pored sinhronizacije, jedna od najčešće posmatranih dinamika na mreži je dinamika slučajnih šetnji. Ova dinamika je poslužila kao osnova za razvoj različitih metoda topoloških karakteristika u mrežama [31, 36, 37]. Svi ovi metodi su bazirani na činjenici da su karakteristike slučajnih šetnji jako korelisane sa strukturom same mreže i cinjenici da šlučajni šetač kada se jednom nadje u modulu u njemu ostaje vrlo dugo vremena. Praćenjem odredjenih veličina, kao što je rastojanje slučajnog šetača od početnog čvora, moguće je odrediti članove zajednica kao i njihov broj.
Sinhronizacija i slučajne šetnje spadaju u difuzne dinamike koje se matematički mogu zapisati preko odgovarajućih matrica, Laplasijana L. Elementi matrice L direktno zavise od elemenata matrice povezanosti, A, odnosno matrice težina, W [38, 28, 39]. U spektru Laplasove matrice, postojanje modula je u vezi sa najmanjim svojstvenim vrednostima u mreži [1, 30, 27]. Lokalizovanost svojstvenih vektora koji odgovaraju najmanjim svojstvenim vrednostima Laplasijana, takodje odražava modularnu strukturu. Spektralni metod baziran na raličitim matricama koje opisuju strukturu mreže se može uspešno primeniti za nalaženje njenih podstruktura [27, 28]. Jedan od nesdostaka metoda je što ne može biti primenjen na usmerene mreže, zbog nesimetrične forme odgovarajućih matirca.
U statističkoj analizi podataka, jedna od standardnih procedura za odredjivanje parametara raspodele iz nekog skupa podataka je *metod maksimalne verodostojnosti* (MMV). Ideja ovog metoda je da se nadje maksimalna vrednost funkcije verodostojnosti za dati skup podataka u odnosu na parametre predpostavljene raspodele da bi se ocenila vrednost tih parametara odnosno sama raspodela. Ukoliko podaci u nekom uzorku potiču iz više različitih raspodela neophodno je primeniti teoriju mešovitih modela da bi se odredio oblik raspodele i zatim primeniti MMV u cilju

odredjivanja rezultujuće raspodele. Kao što je već istaknuto, čvorovi u mrežama mogu biti grupisani prema nekom kriterijumu (u slučaju modula čvorovi u istom modulu imaju iste obrasce povezivanja), što ukazuje na mogućnost primene teorije mešovitih modela i metoda maksimalnih verovatnoća za identifikaciju ovih modularnih struktura. Metod predložen u radu [34] je formulisan za slučaj neotežinjenih usmerenih i neusmerenih grafova. Generalizacija ovog metoda za slučaj otežinjenih mreža [40] omogućava njegovu primenu na veliki skup realnih sistema. Kao što će biti pokazano metod je vrlo efikasan u nalaženju zajednica u mrežama generisanim iz modela. U ovim mrežama ne postoji *a priori* uvedena hijerarhijska struktura zajednica, kao ni prepokrivanje istih zbog čega sposobnost metoda za nalaženje ovakvih struktura nije testirana.
U drugoj glavi ove teze opisan je generalizovani metod za nalaženje podgrafova u otežinjenim mrežama (oMMV) baziran na metodu maksimalne verodostojnosti kao i implementacija metoda. Pokazano je da je metod za nalaženje modula u binarnim grafovim (MMV) [34] samo specijalan slučaj generalizovanog metoda. Treća glava sadrži rezultate testova metoda na usmerenim (neumserenim) otežinjenim (binarnim) modularnim grafovima generisanim iz modela opisanih u radovima [27, 40].
U četvrtoj glavi teze predstavljeni su rezultati primene oMMV za nalaženje modularne strukture nekih realnih mreža. Procedura mapiranja interakcija gena na graf se svodi na nalaženje korelacija izmedju njihovih ekspresija tokom ćelijskog ciklusa. Iz toga sledi da je najpogodniji način za predstavljanje genskih interakcija upravo otežinjeni graf. Iskoristili smo oMMV da bi smo identifikovali module u mreži genskih ekspresija organizma *Saccharomyces cerevisiae*, poznatog kao pivski kvasac.
Metod je primenjen i na neke od mreža tehnološki posredovanih socijalnih interakcija korisnika veb portala kao što su filmska baza IMDb [41] i blog sajta B92 [42]. Podatke u vidu korisnika i filmova odnosno postova i komentara je moguće predstaviti bipartitnim grafom. Analizom projekcija ovih grafova na jedan od tipova čvorova korišćenjem oMMV dobijaju se zajednice koje nam pružaju neophodne informacije o ponašanja korisnika odredjenog veb portala.
Sve slike mreža prikazane u ovoj tezi napravljene su uz pomoć programskog paketa pajek [43]

Glava 2

Metod maksimalne verodostojnosti

Primena metoda maksimalne verodostojnosti (MMV) za identifikaciju modularne strukture u binarnim mrežama opisana je u radu [34]. MMV za binarne mreže predstavlja samo specijalan slučaj opšteg, generalizovanog, metoda (oMMV) [40]. Ideju i sam metod ćemo predstaviti za najopštiji oblik mreže, usmerena otežinjena mreža, a zatim pokazati da je MMV opisan u radu [34] samo poseban slučaj oMMV za grafove sa jediničnim težinama linkova.

2.1 Generalizovani metod

Metod predstavljen u ovom poglavlju je generalizacija metoda maksimalne verodostojnosti koji koristi tehnike *mešovitih modela* i *estimaciono-maksimizacinog algoritma* za nalaženje modula u mrežma. Kao što je već rečeno, otežinjena mreža od N čvorova može biti predstavljena matricom težina W koja se sastoji od N^2 elemenata. Elementi matrice W_{ij} imaju vrednost w ($w \in R$) ukoliko izmedju čvorova i i j postoji link jačine w, odnosno 0 u svim ostalim slučajevima. Težina linka izmedju dva čvora se može posmatrati kao postojanje w linkova jedinične težine, t.j. otežinjeni grafovi postaju multigrafovi [44]. Osnovna predpostavka modela je da je čvorove u mreži moguće podeliti u G grupa prema nekom svojstvu. Pripadnost čvora i nekoj grupi iskazana je veličinom g_i. Vrednosti veličina g_i nisu poznate, zbog čega se iste označavaju kao *skriveni podaci*.
Kao što je već naglašeno, metod se može primeniti kako na usmerene tako i na neusmerene mreže, ali će zbog jednostavnosti prvo biti razmatran slučaj usmerenih mreža. Modul u otežinjenim mrežama predstavlja skup čvorova medjusobno povezanih linkovima čije su težine veće u odnosu na težine linkova izmedju čvorova koji pripadaju raličitim grupama. Iz definicije modula na mreži sledi da čvorovi koji pri-

padaju istoj grupi imaju sličan skup prvih suseda u odnosu na najače linkove koji polaze od njih, odnosno na sličan način "vide" ostatak mreže kroz linkove najvećih težina. Model je karakterisan dvema grupama parametara. Jedna grupu parametara modela, $\{\theta_{ri}\}$, gde $r \in \{1, G\}$ i $i \in \{1, N\}$, se odnosi na način grupisanja čvorova. θ_{ri} predstavlja verovatnoću da postoji link od slučajno odabranog čvora iz grupe r ka čvoru i. Verovatnoća da postoji w linkova od slučajno odabranog čvora iz grupe r ka čvoru i je data kao θ_{ri}^{w}. Drugu grupu parametara, π_r, predstavljaju verovatnoće da slučajno izabran čvor pripada grupi r. Verovatnoće zadovoljavaju uslove normalizacije,

$$\sum_r \pi_r = 1 \ , \quad \sum_i \theta_{ri} = 1 \ . \tag{2.1}$$

Veličine u modelu je moguće svrstati u tri klase: podaci koje merimo, W, podaci koje tražimo, $\{g_i\}$ i parametri modela $\{\theta_{ri}\}$ i $\{\pi_r\}$. Dalje u tekstu, oznaka π će odgovarati grupi parametara $\{\pi_r\}$, θ će odgovarati $\{\theta_{ri}\}$, dok se g odnosi na skup skrivenih podataka $\{g_i\}$.

Parametre je moguće odrediti metodom maksimalne verodostojnosti (oMMV) kombinovane sa estimaciono-maksimizacionim algoritmom. U ovom slučaju problem se svodi na nalaženje maksimalne vrednosti funkcije verodostojnosti $Pr(W, g|\pi, \theta)$ nalaženjem odgovarajuće vrednosti parametara θ i π. Verodostojnost $Pr(W, g|\pi, \theta)$ može biti shvaćen i kao uslovna verovatnoća da se dobije mreža W čija je modularna struktura odredjena vrednostima veličina g za odredjenu vrednost parametara π i θ. Funkciju $Pr(W, g|\pi, \theta)$ je moguće izraziti preko funkcija verodostostojnosti $Pr(W|g, \pi, \theta)$ i $Pr(g|\pi, \theta)$ koristeći pravilo faktorizacije,

$$Pr(W, g|\pi, \theta) = Pr(W|g, \pi, \theta) Pr(g|\pi, \theta) \ . \tag{2.2}$$

$Pr(W|g, \pi, \theta)$ predstavlja uslovnu verovatnoću za realizaciju mreže W ako postoji podela mreže data sa g za parametre π i θ. Da bi smo odredili oblik funkcije $Pr(W|g, \pi, \theta)$ neophodno je posmatrati jedan link koji ide od čvora i ka čvoru j, težine W_{ij}. Verovatnoća za ovaj link je jednaka $\theta_{g_i j}^{W_{ij}}$, gde vrednost g_i odredjuje kojoj grupi pripada čvor i. Verovatnoća za sve linkove u mreži je jednaka

$$Pr(W|g, \pi, \theta) = \prod_{ij} \theta_{g_i j}^{W_{ij}} \ . \tag{2.3}$$

Verodostojnost $Pr(g|\pi, \theta)$ je data jednačinom

$$Pr(g|\pi, \theta) = \prod_i \pi_{g_i} \ . \tag{2.4}$$

Iz jednačine 2.2 sledi,

$$Pr(W, g|\pi, \theta) = \prod_i \pi_{g_i} \prod_j \theta_{g_i, j}^{W_{ij}} \ . \tag{2.5}$$

Metod maksimalne verodostojnosti je metod izbora jedne vrednosti parametara modela kao ocene tih parametara, ali tako da funkcija verodostojnosti 2.5 ima maksimalnu vrednost. Zbog jednostavnosti umesto maksimalne vrednosti funkcije verodostojnosti date jednačinom 2.5 traži se ocena maksimalne vrednosti njenog logaritma

$$L = \sum_i [\ln(\pi_i) + \sum_j W_{ij} \ln(\theta_{g_i j})] \ , \tag{2.6}$$

koji ima maksimum za iste vrednosti parametara π i θ kao i funkcija 2.5.
Kako su veličine g nepoznate, to za posledicu ima da je i vrednost funkcije verodostojnosti, odnosno njen logaritam, takodje nepoznat. Iz tih razloga neophodno je usrednjiti logaritam verodostojnosti, jednačina 2.6, preko distribucije verovatnoća za veličine g, $Pr(g|A, \pi, \theta)$,

$$\begin{aligned} \overline{L} &= \sum_{g_1}^{G} \cdots \sum_{g_n}^{G} Pr(g|W, \pi, \theta) \sum_i [ln\pi_{g_i} + \sum_j W_{ij} ln\theta_{g_i j}] \\ &= \sum_{ir} q_{ir} [ln\pi_r + \sum_j W_{ij} ln\theta_{rj}] \ . \end{aligned} \tag{2.7}$$

Veličina $q_{ir} = Pr(g_i = r|W, \pi, \theta)$ predstavlja verovatnoću da čvor i pripada grupi r i data je jednačinom

$$q_{ir} = Pr(g_i = r|W, \pi, \theta) = \frac{Pr(W, g_i = r|\pi, \theta)}{Pr(W|\pi, \theta)} \ . \tag{2.8}$$

Kao krajnji rezultat postupka maksimizacije funkcije 2.8 dobijaju se upravo verovatnoće q_{ir} a ne brojevi g_i. Imenilac u jednačini 2.8 se može izračunati sumiranjem po svim mogućom vrednostima veličine g u jednačini 2.5:

$$\begin{aligned} Pr(W, g_i = r|\pi, \theta) &= \sum_{i_1=1}^{G} \cdots \sum_{i_N=1}^{G} \delta_{g_i r} Pr(W, g|\pi, \theta) \\ &= \sum_{i_1=1}^{G} \cdots \sum_{i_N=1}^{G} \delta_{g_i r} \prod_k \pi_k \prod_j \theta_{g_k j}^{W_{kj}} \\ &= \pi_r \prod_j \theta_{rj}^{W_{kj}} \prod_{k \neq i} \sum_{s=1}^{G} \pi_s \prod_j \theta_{rj}^{W_{kj}} \ , \end{aligned} \tag{2.9}$$

dok je brojilac u jednačini 2.8 jednak

$$\begin{aligned} Pr(W|\pi,\theta) &= \sum_{i_1=1}^{G}\dots\sum_{i_N=1}^{G} Pr(W,g|\pi,\theta) \\ &= \sum_{i_1=1}^{G}\dots\sum_{i_N=1}^{G}\prod_k \pi_k \prod_j \theta_{g_k j}^{W_{kj}} \\ &= \prod_k \sum_{s=1}^{G} \pi_s \prod_j \theta_{rj}^{W_{kj}} \ . \end{aligned} \tag{2.10}$$

Na osnovu jednačina 2.10 i 2.11 sledi

$$q_{ir} = \frac{\pi_r \prod_j \theta_{rj}^{W_{ij}}}{\sum_s \pi_s \prod_j \theta_{sj}^{W_{ij}}} \ . \tag{2.11}$$

Maksimalna vrednost $\overline{L}$ predstavlja najbolju procenu maksimalne vrednosti L, dok odgovarajuća procena vrednosti parametara π i θ najbolju ocenu položaja maksimuma u prostoru parametara.
Maksimalnu vrednost funkcije 2.8 je moguće naći analitičkim putem korišćenjem metoda Lagranžovih množitelja da bi se uključili i normalizacioni uslovi dati jednačinom 2.1. Lagranžova funkcija je data

$$F = \sum_{ks} q_{ks}[\ln \pi_s + \sum_j W_{kj} \ln \theta_{sj}] + \lambda(\sum_s \pi_s - 1) + \sum_s \mu_s(\sum_k \theta_{sk} - 1) \ , \tag{2.12}$$

gde $\mu = \{\mu_1, \dots, \mu_s\}$ označava set Lagranžovih množitelja uz jednačine $\sum_i \theta_{si} - 1 = 0$. Iz uslova $\frac{\partial F}{\partial \lambda} = 0$ i $\frac{\partial F}{\partial \mu_s} = 0$ dobijaju se normalizacioni uslovi za parametre π i θ. Parametri π i θ izraženi kao funkcije verovatnoća q_{ir} dobijaju iz uslova $\frac{\partial F}{\partial \pi_r} = 0$

$$\pi_r = \frac{1}{\lambda} \sum_k q_{kr} \ , \tag{2.13}$$

odnosno $\frac{\partial F}{\partial \theta_{ri}} = 0$

$$\theta_{ri} = \frac{1}{\mu_r} \sum_k q_{kr} W_{ki} \ , \tag{2.14}$$

Vrednost Lagranžovih množitelja jednostavno je odrediti iz uslova normalizacije 2.1. Relacije za parametre π i θ izražene preko verovatnoća q_{ir} date su:

$$\pi_r = \frac{\sum_i q_{ir}}{n} \ , \quad \theta_{ri} = \frac{\sum_j A_{ji} q_{jr}}{\sum_j l_j q_{jr}} \ . \tag{2.15}$$

Ovde $l_i = \sum_k W_{ki}$ predstvalja *jačinu čvora* u odnosu na linkove koji polaze od njega, *izlazeće linkove*.
Izloženi metod se odnose na slučaj otežinjenih usmerenih grafova, medjutim većina izučavanih mreža je neusmerena. Predstavljeni metod nije pogodan za slučaj neusmerenih mreža zbog nesimetričnog odnosa izmedju čvorova u mreži, t.j. čvorovi se grupisu prema linkovima koji polaze od njih, nezavisno od toga kako su njihovi susedi povezani. U neusmerenim mrežama postoji simetrija linkova. Da bi gore izložen metod bio pogodan za nalaženje modularne strukture u neusmerenim mrežama neophodno je uvrstiti simetriju veza. Ponovo definišmo parametar θ_{ri} kao verovatnoću da je slučajno izabran čvor iz grupe r povezan za čvorom i, medjutim sada moram uvesti i uslov da se link formira samo ako oba čvora odaberu jedan drugog. Verovatnoća da postoji link izmedju čvorova i i j je data kao $\theta_{ri}\theta_{sj}$ gde je $r(s)$ grupa kojoj pripada čvor $i(j)$. Ova verovatnoća zadovoljava uslov normalizacije $\sum_{ij} \theta_{ri}\theta_{sj} = 1$ za svako r i s dok za $r = s$ dobijamo

$$\sum_{ij} \theta_{ri}\theta_{rj} = [\sum_{ij} \theta_{ri}]^2 = 1 \ . \tag{2.16}$$

Verovatnoća $Pr(W|g, \pi, \theta)$ je data

$$Pr(W|g, \pi, \theta) = \sum_{i>j} [\theta_{g_i j}\theta_{g_j i}]^{W_{ij}} = \sum_{ij} \theta_{g_i j}^{W_{ij}} \ , \tag{2.17}$$

gde je iskorišćena činjenica da W_{ij} ima simetričan oblik, $W_{ij} = W_{ji}$.
Ostala izvodjenja, rezultati, kao i implementacija algoritma su isti kao i za slučaj usmerenih grafova.
Iteracijom jednačina 2.11 i 2.15 moguće je oceniti vrednosti parametara za koje očekivana vrednost logaritma funkcije verodostojnosti ima maksimalnu vrednosti i naći verovatnoće pripadnosti čvorova grupama. Numerička implementacija je opisana u poslednjem odeljku ove glave.

2.1.1 Metod maksimalne verodostojnosti za binarne grafove

Binarni grafovi su samo specijalni slučaj otežnjenih grafova kod kojih svi linkovi imaju težine jednake 1. Matrica težina tada postaje identična matrici povezanosti A. Sada se podgraf ili modul definiše kao skup čvorova koji imaju isti (ili sličan) broj prvih suseda preko svih svojih odlazećih linkova u slučaju usmerene mreže. U slučaju neusmerene mreže svi linkovi se mogu posmatrati kao odlazeći. Kao i u slučaju otežinjenih grafova, veličinama g_i je označena grupna pripadnost čvorova, dok su parametri modela verovatnoće π i θ. Funkcija maksimalne verodostojnosti za neoteženje mreze reprezentovane matricom A data je kao

$$Pr(A, g|\pi, \theta) = \prod_i \pi_{g_i} \prod_j \theta_{g_i,j}^{A_{ij}} \ . \tag{2.18}$$

Jednostavnom smenom A umesto W u jednačine q_{ir}, θ_{ir} i π_r dobijau se jednačine za verovatnoće q_{ir}

$$q_{ir} = \frac{\pi_r \prod_j \theta_{rj}^{A_{ij}}}{\sum_s \pi_s \prod_j \theta_{sj}^{A_{ij}}} , \tag{2.19}$$

i parametare modela

$$\pi_r = \frac{\sum_i q_{ir}}{n} , \quad \theta_{ri} = \frac{\sum_j A_{ji} q_{jr}}{\sum_j q_{out}(j) q_{jr}} . \tag{2.20}$$

Jačina čvora u jednačini 2.11, za slučaj binarnih grafova postaje izlazni stepen čvora, jednačina 2.19. Dobijene jednačine za verovatnoće q_{ir} i parametre π i θ se slažu sa jednačinama koje su originalno izvedene za binarnih graf u referenci [34].

2.2 Numerička implementacija algoritma

Maksimum očekivane vrednosti funkcije verodostojnosti moguće je naći primenom estimaciono-maksimizacionog algoritma. Algoritam se sastoji u sukcesivnoj iteraciji estimacionih i maksimizacionih koraka, odnosno jednačina 2.11 i 2.15 za otežinjene grafove ili jednačina 2.19 i 2.20 za slučaj binarnih mreža. U estimacionom koraku se izračunava očekivana vrednost funkcije verodostojnosti oderdjivanjem vrednosti parametara q_{ir} u odnosu na trenutnu vrednost parametara modela, dok maksimizacioni korak predstavlja odredjivanje vrednosti parametara modela koje maksimizuju očekivanu vrednost logaritma funkcije verodostojnosti.
Prvo je neophodno zadati početne vrednosti parametrima modela θ i π pazeći na normlizacioni uslov dat jednačinom 2.1. Izbor inicijalnih vrednosti parametara može biti različit, od slučajnog do vrednosti parametara zadatih iz podele mreže dobijene nekom od komplementarnih metoda. U implementaciji algoritma testirana su dva načina inicijalizacije parametara modela: inicijalizacija slučajnim izborom iz uniformne razspodele i iz Gausove raspodele. Uniformni odabir parametara π_r ili θ_{ri} podrazumeva pridruživanje unformno odabranog broja iz intervala $[0, 1]$ a zatim i normalizovanje parametara tako da su zadovoljeni uslove 2.1. Drugi način odabira je iz Gausove distribucije. Sistem jednačina 2.11 i 2.15 odnosno 2.19 i 2.20 ima jednu trivijalnu fiksnu tačku koja je zadata vrednostima parametara $\pi_r = \frac{1}{G}$ i $\theta_{ri} = \frac{1}{N}$ za svako r i i. Ovde treba istaći da je broj grupa G ulazni parametar, odnosno neophodno ga je odrediti nekim drugim metodom. Ukoliko je inicjalna vrednost parametara zadata vrednostima ove fiksne tačke iteracija jednačina za vrovatnoće q_{ir} i parametre modela ne vodi ka njihovoj novoj vrednosti odnosno novoj podeli mreže. Nasuprot tome, ukoliko početne vrednosti parametara malo odstupaju od onih koje odgovaraju fiksnoj tački, iteracijom jednačina 2.11 i 2.15 odnosno 2.19 i 2.20 sistem će konvergirati ka nekom od lokalnih maksimuma očekivane vrednosti

funkcije vredostojnosti i samim tim nekoj novoj vrednosti parametara modela i verovatnoća q_{ir} [40, 34]. Početne vrednosti je moguće zadati iz Gausijana sa srednjom vrednošću $\frac{1}{G}$ (za parametre π) odnosno $\frac{1}{N}$ (za paramtere θ), i standardnom devijacijom $0 < \varepsilon << 1$. Normiranjem dobijenih vrednosti parametara modela tako da zadovoljavaju jdenačinu 2.1 zadajemo poečetnu podelu mreže na grupe. Iz testova MMV i oMMV metoda na mrevžama sa poznatom modularnom strukturom, sledi da ukoliko se početne vrednosti parametara zadaju Gausovim vreovatnoćama algoritam brže konvergira nego u slučaju parametara zadatih unformnom raspodelom. U nastavku teze podrazumeva se izbor početnih vrednosti parametara iz Gausove distribucije.

Iteracijom jednačina za q_{ir} odnosno π i θ oredjuje se njihovo samousaglašeno rešenje koje odgovara jednom od lokalnih maksimuma funkcije verodostojnosti. Numerička implementacija algoritma zahteva definisanje *kontrolnih parametara* kojima se kontroliše da li je algoritam konvergirao ka fiksnoj tački. Smatraćemo da je algoritam konvergirao ukoliko su ispunjeni sledeći uslovi: $|\pi^(o)_r - \pi_r| < \varepsilon_\pi$ odnosno $|\theta^(o)_{ri} - \theta_{ri}| < \varepsilon_\theta$ za svako i i r, gde π^o i θ^o predstavljaju vrednost parametara u predhofnom iteracionom koraku. Sve podele mreže u ovoj tezi su nadjene za vrednosti kontrolnih parametara $\varepsilon_\pi = \varepsilon_\theta = 10^{-8}$.

Na osnovu početnih vrednosti parametara modela moguće je izračunati početne vrednosti verovatnoća q_{ir} i samim tim odrediti početnu podelu mreže na G grupa. Svakom iteracijom jednačina 2.11 i 2.15 odnosno 2.19 i 2.20, dobija se nova podela mreže čija funkcija verodostojnosti ima veću očekivanu vrednost nego predhodna podela. Posle odredjenog broja iteracija, estimaciono-maksimizacioni algoritam konvergira ka lokalnom maksimumu očekivane vrednosti funkcije verodostojnosti dajući procenjene vrednosti verovatnoća q_{ir} za zadate početne vrednosti parametara modela i kao rezultat modela dobija verovatnoće q_{ir} iz kojih se zatim odredjuje pripadnost čvora odredjenoj grupi. U najvećem broju slučajeva algoritam čvor i pridruži nekoj grupi r' sa verovatnoćom $q_{ir'} = 1$, odakle sledi $g_i = r^{'}$. Ukoliko to nije slučaj, pripadnost grupama je moguće odrediti na sledeći način: ukoliko je $q_{ir'} > 0.5$ tada čvor i pripada grupi r' i $g_i = r'$. U nekim slučajevima algoritam nije u mogućnosti da neke čvorove pridruži nekoj od grupa, sve verovatnoće su $q_{ir} < 0.5$, tada se kaže da grupu čvora nije mogué odrediti.

Većina funkcija definisanih na kompleksnim mrežama ima veliki broj maksimuma (odnosno minimuma) što čini nalaženje njihovih ekstremalnih vrednosti zahtevnim i kompleksnim zadatkom. Funkcija verodostojnosti, odnosno njen logratiam, takodje imaju više lokalnih maksimuma zbog čega nadjena podela mreže ne mora biti idealna. U cilju nalaženja idealne podele neophodno je primeniti metod na istoj mreži više puta za različite inicjalne vrednosti parametara i od dobijenih podela odabrati onu sa najvećom vrednošću maksimuma verodostojnosti. Svi rezultati u tezi su dobijeni izvršavanjem algoritma 100 puta a zatim odabirom podele koja ima najveći maksimum.

Programski kod numeričke implementacije metoda koji se može primeniti na sva četiri opisana tip mreža je dat u dodatku A.

Glava 3

Modularna struktura kompjuterski generisanih grafova

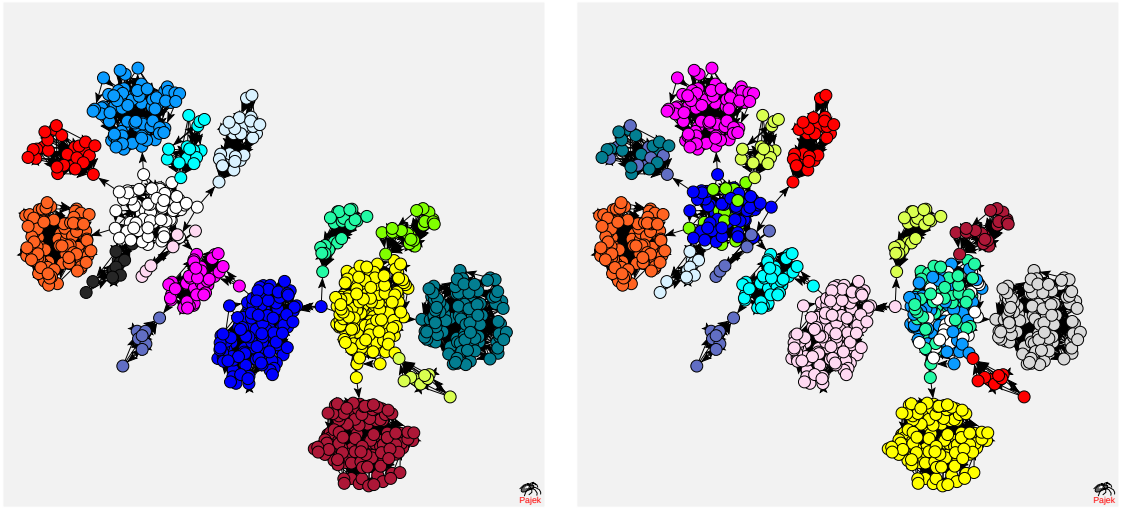
Za testiranjanje efikasnosti i preciznosti MMV i oMMV metoda neophodni su modeli mreža čije su osobine dobro poznate i mogu se lako varirati promenom vrednosti odgovarajućih parametara. U ovoj glavi biće predstavljeni rezultati metoda primenjenih na dva modela kompjuterski generisanih mreža sa dobro kontrolisanom strukturom. Modelom modularnog binarnog grafa [40, 27] moguće je generisati usmerene kompleksne mreže sa modularnom strukturom, gde su veličina i broj modula statističke varijable. Raspodela povezanosti mreža generisanih ovim modelom je stepenog oblika što sam model čini mnogo realističnijim u odnosu na modele modularnih mreža koji se standardno koriste. Primenom MMV na mreže generisane ovim modelom biće pokazano kako njegova efikasnost zavisi od lokalnih i globalnih osobina mreže kao i izbora vrednosti ulaznog parametra G.
Najjednostavniji model za generisanje mreža je Erdoš-Renji (ER) model [8] kao i njegove raličite varijacije. Iako se grafovi generisani različitim verzijama ER modela u mnogome razlikuju od realnih mreža upravo zbog svoje jednostavnosti su vrlo pogodni za testiranje različitih metoda. Za testiranje oMMV koristićemo otežinjeni Erdoš-Renji graf sa modularnom strukturom objavljen u radu [40].

3.1 Binarni modularni grafovi i MMV

U ovom poglavlju biće predstavljeni rezultati primene MMV na grafove generisane modelom modularnih usmerenih mreža [40, 27]. Kao osnova za model poslužio je model klasterisanih usmerenih *scale free* mreža objavljen u radu [11]. Preferencijalno povezivanje čvorova i preferencijalno prevezivanje linkova u toku rasta mreže za rezultat imaju korelisani graf sa statističkim osobinama sličnim WWW [11]. Netrivijalnom generalizacijom modela, uvodjenjem pravila da se sa verovatnoćom P_0 u

svakom trenutku započinje sa formiranjem novog modula, dobija se model modularnih grafova [40, 27].
Pravila rasta mreže ovim modelom kao i osobine mreža data su u referencama [40, 27]. Ovde će samo ukratko biti opisano kakva se struktura mreže može očekivati za pojedine vrednosti parametara. Struktura mreža je kontrolisana sa tri parametra: prametrom α se kontroliše prevezivanje linkova, parametar M kontroliše gustinu linkova u mreži dok P_0 srednji broj podgrafova. Za vrednosti $P_0 = 0$ dobijaju se klasterisane mreže bez modularne strukture čije osobine zavise od vrednosti parametra M i α. Za $M = 1$ i $\alpha = 1$ generisani graf je *scale free* drvo, dok se pojavljivanje cikličnih putanja javlja za $\alpha < 1$ odnosno kada se dozvoli prevezivanje odredjene frakcije, $(1 - \alpha)$, linkova. Povećanjem parametra M dobijaju se korelisane mreže čija klasterisanost opet zavisi od parametra α, što je manja vrednost parametra α to je graf više klasterisan. Pojava modula u mreži je moguća za vrednosti parametra $P_0 > 0$ i njihov broj je u srednjem jednak NP_0. Struktura modula zavisi od vrednosti parametara M i α na sličan način kao i kod grafova za $P_0 = 0$. Jačina povezanosti čvorova u modulu u odnosu na medjusobnu povezanost samih modula je kontrolisana parametrom α. Za $\alpha = 1$ moduli su nekorelisani povezani grafovi, t.j. izmedju njih postoji samo po jedan usmereni link. Smanjivanjem vrednosti α za fiksiranu vrednost M povećava se broj linkova izmedju modula, odnosno *mreža modula* više nije drvo. Neke od mreža generisane iz modela za raličite vrednosti parametara date su na slikama 3.1, 3.2, 3.3 i 3.4.
Metod MMV je testiran na usmerenim mrežama generisanim opisanim modelom za različite vrednosti parametara α, M i P_0. Za modularne usmerene mreže, $P_0 > 0$, efikasnost i preciznost metoda zavisi od jačine povezanosti čvorova unutar modula kao i broja linkova izmedju podgrafova. Što je veći broj linkova izmedju čvorova u odredjenoj zajednici to je veći broj onih "vide" ostatak mreže na sličan način što garantuje da će ih metod svrstati u istu grupu dok jasnoća granice izmedju modula zavisi od broja prevezanih linkova. Što je veći broj linkova izmedju modula (manja vrednost parametra α) to je različivost modula manja pa samim tim i podudaranje podele dobijene MMV sa originalnom. U ovom radu smo odabrali par primera karakterističnih grafova kao ilustraciju primene metoda.

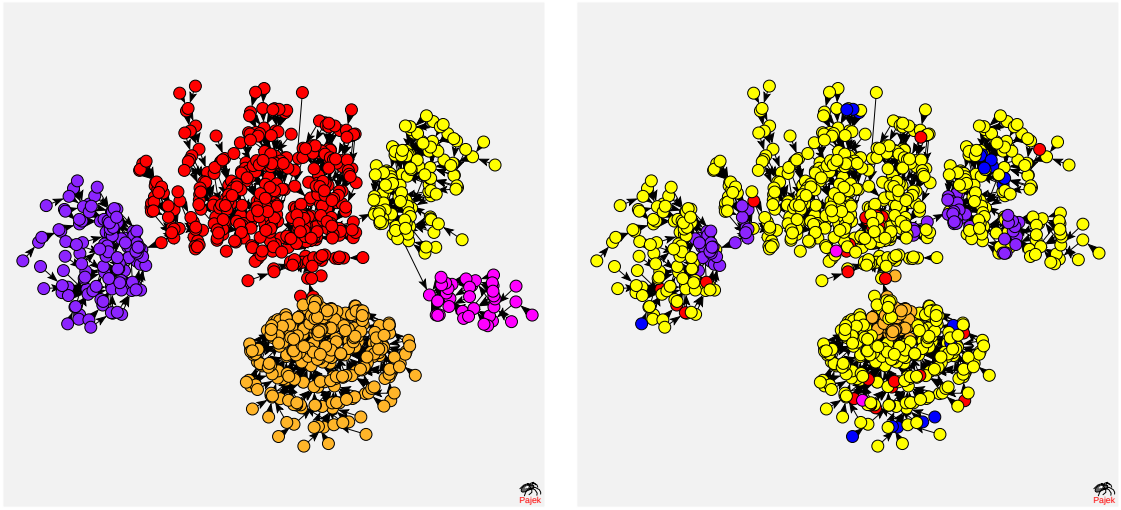
Mreže generisane iz modela za vrednosti parametara $\alpha = 1$, $M > 1$ i $P_0 > 0$ su ekstremni primer grafova sa modularnom strukturom, 3.1 (levo). Granica izmedju grupa čvorova je u ovim grafovima vrlo jasna, i kao takvi predstavljaju dobar primer za testiranje različitih metoda za nalaženje podgrafova u mrežama. U strukturi svakog od podgrafova se mogu razlikovati centralni čvorovi, odlikuju se velikim brojem ulaznih linkova u poredjenju sa ostalima, za koje su povezani skoro svi ostali čvorovi u modulu. Centralni čvorovi imaju vrlo bitnu ulogu u mreži, upravo preko njih su povezani moduli, što je direktna posledica pravila modela [27]. Primena MMV na ovim grafovima sa različitim brojem modula pokazuje da efikasnost metoda zavisi od broja modula u mreži kao i od veličine mreže. Što je veličina mreže veća



Slika 3.1: Modularna mreža veličine $N = 1000$ čvorova sa srednjim brojem linkova po čvoru $M = 3$, $G_o = 17$ modula ($P_0 = 0.009$) i $\alpha = 1$. Levo je prikazana originalna mreža dok je na desnoj slici prikazana podela grafa odredjena MMV za $G = 17$. Čvorovi iste boje pripadaju istom podgrafu.

to je metod sporiji (duže vreme konvergencije i veći broj mogućih maksimuma) a podele dobijene metodom sve više odstupaju od originalnih koje slede iz modela. Iz testova sledi da je optimalna veličina mreže do dve hiljade čvorova sa brojem modula manjim od $G = 10$. Na slici, 3.1 (desno) je prikazana podela mreže koja se sastoji od 17 modula ($P_0 = 0.009$) sa srednjim brojem linkova po čvoru $M = 3$ nadjena MMV za usmerene grafove. Postoji izvesna tendencija metoda da više malih grupa od po svega par čvorova svrstava u zajedničku grupu ili grupe, a da veće module deli na podgrafove. Podela većih modula je posledica načina na koji metod grupiše čvorove. Veći moduli mogu sadržati podgrupe čvorova koje se razliku prema obrascima povezivanja, što za posledicu ima da se kao pozicija maksimuma funkcije verodostojnosti javlja set vrednosti parametara θ i π koji ne odgovara originalnoj podeli.

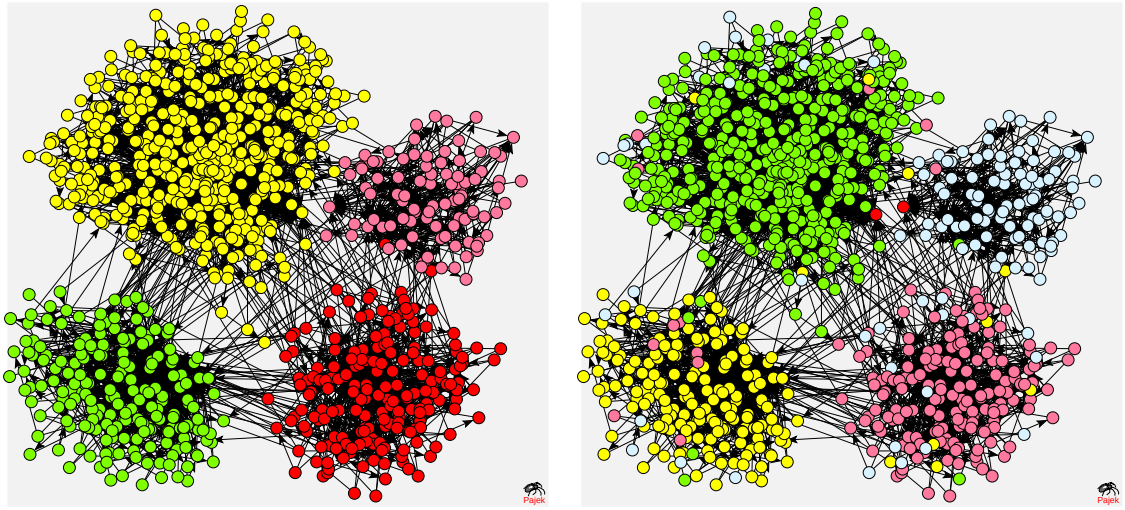
Drugi primer mreža na kojima je testiran metod su grafovi kod kojih moduli imaju strukturu drveta i poznati su kao *drvo drveta* [27]. Iako grupe čvorova ovde ne predstavljaju module u smislu date definicije, postoje metodi, na primer spektralni metod [27], pomoću kojih se mogu identifikovati aciklični podgrafovi u mreži. Uporedjivanje rezultata MMV za $G = 5$ sa originalnom podelom mreže sa slike 3.2 (levo) pokazuje da metod nije pogodan za nalaženje modula sa strukturom drveta u mreži. U drvetu svaki čvor ima po jednog prvog suseda u odnosu na svoj odlazeći link i metod je uspešan u nalaženju ovih grupa 3.2 (desno). Medjutim metod nema definisan kriterijum po kom će nadjene podgrupe grupisati tako da odgovaraju



Slika 3.2: Podgrafovi na drvetu-drveta ($M = 1$, $\alpha = 1$, $P_0 = 0.005$) nadjeni MMV metodom za $G = 5$ koji je jednak originalnom broju podgrafova (drveta).

originalu, žbog čega metod nije pogodan za traženje modula sa strukturom drveta. Primena MMV na drveta pokazuje mnogo sporiju konvergenciju (potreban je veći broj koraka) estimaciono-maksimizacionog algoritma u odnosu na mreže sa većim M. Drvo predstavlja poseban slučaj bipartitnih grafova. U bipartitnim grafovima čvorovi se mogu podeliti u dve grupe tako da su dozvoljeni linkovi samo izmedju čvorova koji pripadaju različitim particijama. U radu [34] je pokazano da se MMV za $G = 2$ mogu uspešno identifikovati particije u cikličnim binarnim grafovima. Kao što je opisano u predhodnom paragrafu, metod se ne može iskoristiti za nalaženje particija u acikličnim bipartitnim mrežama zbog malog broja linkova izmedju čvorova ($M = 1$) u ovim grafovima.

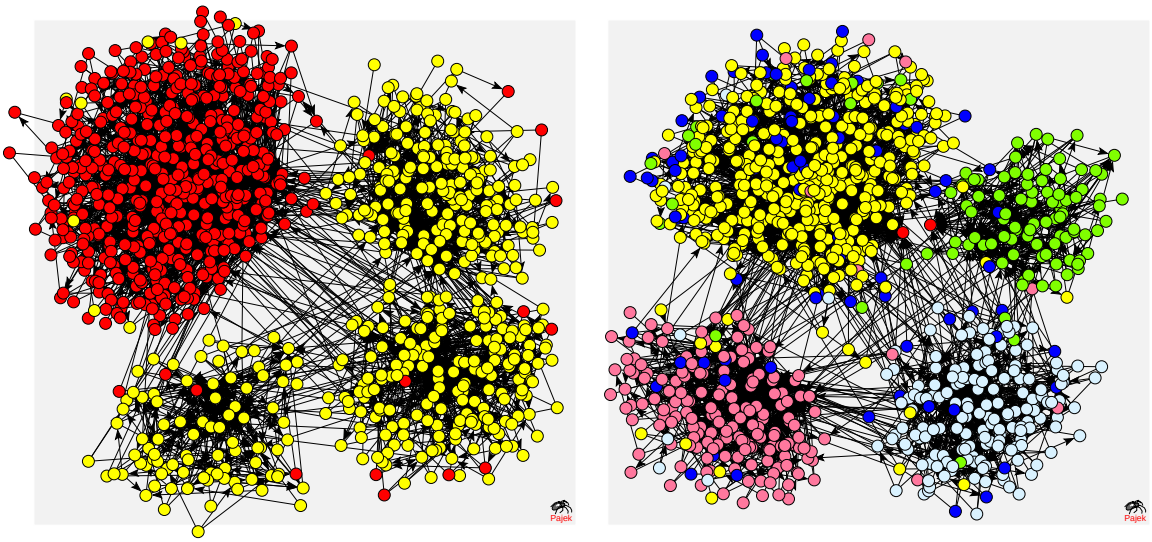
Većina realnih mreža ima modulrnu strukturu gde moduli odgovaraju grupama jače povezanih čvorova, t.j. moduli nemaju strukturu drveta. Medjutim različivost ovih grupa nije jasna, odnosno broj linkova izmedju čvorova u različitim grupama je veći od 1. Iz tih razloga mreže sa $M > 1$ i $\alpha < 1$ predstavljaju dobre kandidate za testiranje metoda maksimalne verodostojnosti. Rezultati primene MMV na mreže generisane za različite vrednosti parametara pokazuju da kao i u slučaju mreža sa $\alpha = 1$ konvergencija estimaciono-maksimizacionog algoritma zavisi od gustine linkova u mreži, M. Preciznost algoritma opada sa povećanjem broja linkova izmedju čvorova u različitim podgrafovima, odnosno smanjivanjem vrednosti parametra α. Smanjivanjem parametra α smanjuje se i podskup čvorova u modulu koji na isti način "vide" ostatak mreže, pa samim tim metod nije u stanju da čvorove grupiše u skaldu sa originalnom podelom. Kao primer uzeta je mrža koja se sastoji od $G_o = 4$ modula ($P_0 = 0.003$), sa srednjim brojem linkova po čvoru $M = 3$ od kojih je 10%



Slika 3.3: (levo) Podela mreže na podgrafove nadjene metodom maksimalne verodostojnosti za $G = 2$ na modularnoj mrež sa 10% prevezanih linkova, srednjom povezanošću $M = 3$ i originalnim brojem grupa $G_o = 4$ ($P_0 0.006$). (desno) Podela mreze nadjena MMV za $G = 4$.

($\alpha = 0.9$) prevezano. Rezultati za $G = 4$, slika 3.3 (desno), pokazuju da je većina čvorova grupisana u skladu sa originalnom podelom, odnosno da je svega $6,8\%$ čvorova svrstano u *pogrešnu* grupu. Neke od čvorova metod nije svrstao ni u jednu grupu (čvorovi crvene boje), t.j. verovatnoće q_{ir} za svako r su manje od 0.5. Upravo ovi čvorovi imaju ulogu konektora u mreži, odnosno oni vide mrežu na potpuno drugačiji način u odnosu na sve ostale čvorove zbog čega ih nije moguće svrstati ni u jednu od grupa.

Kako je G ulazni parametar neophodno je analizirati rezultate metoda za slučajeve kada je vrednost parametra G različita od broja modula u mreži. Ukoliko je $G < G_o$ algoritam neke od grupa spoji odnosno čvorovi koji originalno pripadaju različitim grupama su svrstani u jedna veću grupu. Ovo se jasno vidi na primeru mreže sa $G_o = 4$ modula. Jedan od dva podgrafa nadjena metodom maksimalne verodostojnosti ($G = 2$) predstavlja uniju čvorova iz tri manja originalna modula dok drugi podgraf uglavnom sačinjavaju čvorovi koji pripadaju najvećem modulu u originalnoj podeli, slika 3.4(levo). Za vrednost parametra $G > G_o$ rezultati pokazuju da metod deli veće module na podmodule dok manji moduli ostaju netaknuti ako se izuzme par čvorova, slika 3.4. Iz podela nadjenih sledi da centralni čvorovi, konektori (obojeni crvenom bojom na slikama 3.3 i 3.4 (desno)), nisu svrstani u grupe za $G \geq G_o$, dok su u podelama nadjenim za vrednosti parametra G manjim u odnosu na pravi broj grupa ovi čvorovi dodeljeni jednom od podgrafova, 3.4. Osobina funkcije verodosto-

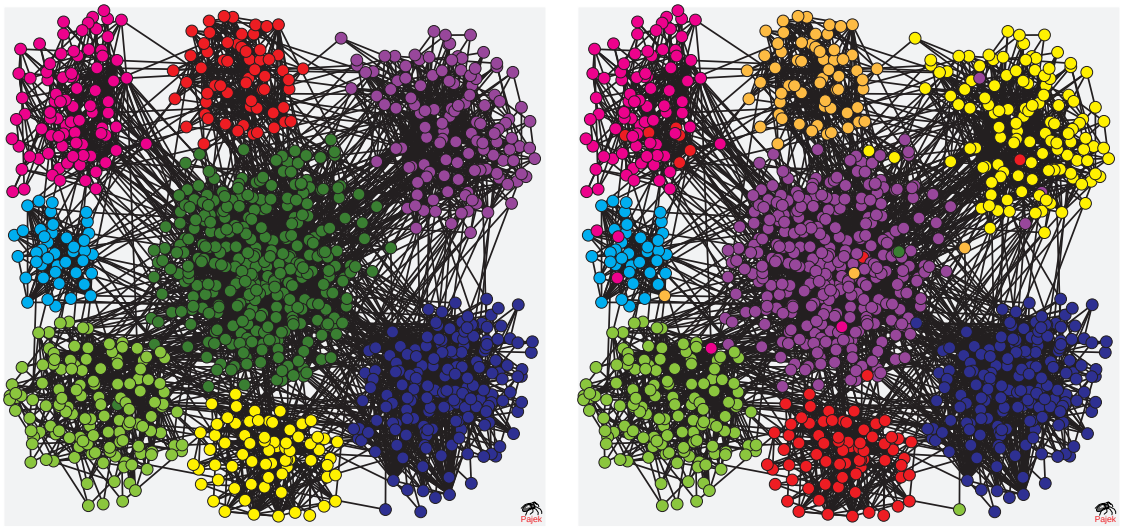


Slika 3.4: Podela mreže prikazane na slici 3.3 nadjena MMV za $G = 2$ (levo) i $G = 5$ (desno).

jnosti je da ona monotono raste sa porastom parametra G. Iz tih razloga ne postoji način da se prostim variranjem vrednosti parametra G oceni idealni broj modula u mreži, već je za ovu procenu neophodno koristi neki drugi metod za nalaženje podgrafova, na primer metod spektralne analize [27].

Neusmerene grafove moguće je generisati istim modelom tako što se jednostavno zanemari smer linka. Ukoliko postoji link izmedju dva čvora u bilo kom smeru u usmerenoj mreži ta dva čvora će biti povezana linkom u neusmerenoj. Primer neusmerene modularne mreže dat je na slici 3.5. Dobijene mreže takodje imaju modularnu strukturu datu iz modela kao i u slučaju usmerenih mreža i pogodne su za testiranje efikasnosti MMV u nalaženju modula u neusmerenim grafovima. Osobine metoda su slične kao i u slučaju usmerenih mreža, metod konvergira brže ukoliko je mreža gušće povezana, dok preciznost metoda opada sa povećanjem broja prevezanih linkova. Poredjenje rezultata metoda za usmerenu mrežu i njenu simetričnu formu pokazuje da je MMV precizniji u nalaženju zajednica u neusmerenoj mreži. Veća preciznosti metoda leži u činjenici da su čvorovi u istom modulu u slučaju neusmerene mreže jače spregnuti, odnosno postoji veće preklapanje izmedju podskupova njihovih prvih suseda.

Kao ilustraciju primenili smo metod na mrežu veličine $N = 1000$ čvorova koja se sastoji od $G_o = 8$ modula sa $M = 4$ linka po čvoru od kojih je 20% ($\alpha = 0.8$) prevezanih, slika 3.5 (levo). Ako se grupe čvorova identifikovane korišćenjem algoritma za $G = 8$, slika 3.5(desno), uporede sa orginalnim koje slede iz modela dobija se da je metod 91% čvorova svrstao u grupu kojoj pripadaju, pri čemu je svaki čvor dodeljen nekoj od grupa. Činjenica da ne postoje čvorovi za koje nije odredjena grupa je



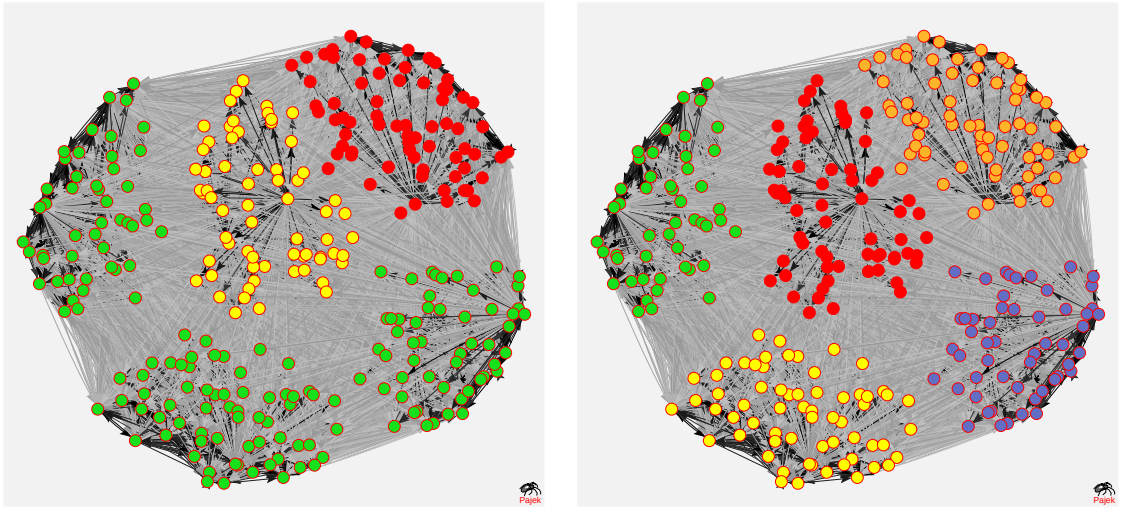
Slika 3.5: Neusmerena modularna mreža sastavljena od $G = 8$ modula sa 20% prevezanih linkova i sredjnom povezanošću $M = 4$. Levo je prikazana originalna podela mreže a desno podela dobijena MMV algoritmom.

direktna posledica neusmerenosti linkova. Čvorovi konektori u slučaju neusmerenih mreža imaju slične prve susede kao i ostatak mreže zbog čega su i svrstani u odgovarajući modul.

3.2 Otežinjeni slučajni grafovi

U drugoj glavi ove teze kao jedan od metoda za nalaženje modularne strukture u realnim otežinjenim mrežama predstavljen je otežinjeni metod maksimalne verodostojnosti. Rezultati primene ovog metoda na mrežama generisanim iz realnih podataka biće opisani u narednoj glavi dok će ovde biti opisane neke od osobina metoda dobijene njegovom primenom na otežinjene mreže sa poznatom modularnom strukturom.

Kao model usmerene modularne mreže sa otežnjenim linkovima biće korišćena jednu od mnogih varijacija Erdoš-Renji grafova opisana u radu [40]. Struktura običnog ER grafa zavisi od parametra p koji predstavlja verovatnoću da postoji link izmedju dva slučajno odabrana čvora. Ovaj parametar kontroliše gustinu povezanosti grafa odnosno srednji stepen čvora. U ER modelu, koji predstavlja statički model grafova, se polazi od N nepovezanih čvorova, a zatim se izmedju svaka dva čvora formira link sa verovatnoćom p. U slučaju usmerenih grafova link koji ide od čvora k ka čvoru j se formira nezavisno od linka $j \to k$. U neusmerenim grafovima formira se samo jedan link od $k \leftrightarrow j$. ER grafovi se nazivaju još i Poasonovim grafovima zbog



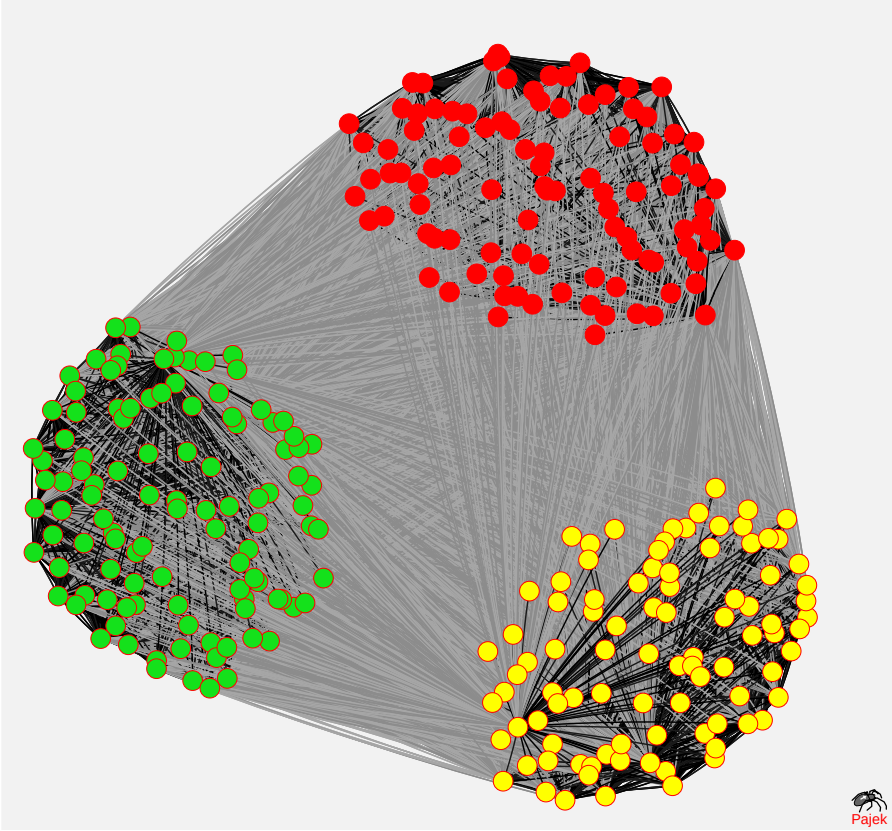
Slika 3.6: Gusto povezani ($p = 0.5$) slučajni modularni graf veličine $N = 300$ čvorova sa otežinjenim usmerenim linkovima gde su težine linkova izmedju čvorova u istoj grupi u intervalu $[12, 18]$ a izmedju čvorova iz različitih grupa u intervalu $[2, 8]$. Na slikama su prikazane podele mreže nadjene primenom oMMV za $G = 3$ (levo) i $G = 5$ desno. Grupe čvorova na slikama predstavljaju originalne grupe odredjene iz modela.

njihove distribucije povezanosti koja je Poasonovog tipa. Ovo znači da za razliku od *scale free* mreža kod kojih postoji visok stepen nehomogenosti čvorova u odnosu na broj prvih suseda, u ER grafovima većina čvorova ima stepen jednak srednjoj vrednosti $< q >$. Ova homogenost je direktno posledica statičke prirode grafa i jednakih verovatnoća formiranja linkova. Otežinjeni ER grafovi se generišu tako što se postojećim linkovima definiše dodatno svojstvo, težina w, koja se bira iz neke raspodele. Ukolik ova raspodela ima oblik stepenog zakona generisani, ER graf postaje nehomogen u odnosu na *jačine čvorova* l_i. Raspodele težina u mrežama koje će biti predstavljene u ovom poglavlju su iz uniformne raspodele diskretnih promenljivih iz skupa $\{a, a + 1, \ldots, b\}$ gde vrednosti granica a i b zavise od konkretne realizacije mreže.

Pored homogene strukture u smislu stepena čvorova ER grafovi su homogeni i na mezoskopskom nivou, t.j. grafovi nemaju modularnu strukturu. Standardni način za generisanje podstruktura u ER grafovima je da se N čvorova unapred podele u G_o grupa a zatim se kreiraju linkovi izmedju čvorova korićenjem dva parametara p_1 i p_2 $(p_2 < p_1)$ [1]. Ukoliko su dva odabrana čvora iz iste grupe, verovatnoća da postoji link izmedju njih je jednaka p_1, dok je verovatnoća da link povezuje dva čvora iz različitih grupa jednaka p_2. Kako je $p_2 < p_1$ to su čvorovi unutar iste grupe gušće povezani odnosno formiraju se topološki moduli u smislu definicije date

u predhodnoj glavi. Različivost modula u mreži zavisi od odnosa $\frac{p_2}{p_1} < 1$, što je njegova vrednost bliža 1 to je struktura mreže homogenija.

U većini relnih otežinjenih mreža modularna struktura je posledica težina linkova



Slika 3.7: Gusto povezani ($p = 0.7$) slučajni modularni graf veličine $N = 300$ čvorova sa otežinjenim neusmerenim linkovima gde su težine linkova izmedju čvorova u istoj grupi u intervalu $[3, 5]$ a izmedju čvorova iz različitih grupa u intervalu $[9, 11]$. Na slikama je prikazana podela mreže nadjene primenom oMMV za $G = 3$. Grupisani čvorovi na slici predstavljaju originalne podgrafove koji slede iz modela.

izmedju čvorova. Ove mreže, kao na primer korelacione matrice gena, su neretko gusto povezane a moduli predstavljaju grupe čvorova koji su medjusobno povezani linkovima relativno većih težina u poredjenju sa težinama linkova koje ti isti čvorovi obrazuju sa ostatkom mreže. Da bi se metod testirao na mrežama sa ovakvom modularnom strukturom neophodna je konceptualno drugačija generalizacija ER modela od one opisane u predhodnom paragrafu. Pravila generazliovanog ER modela kojim se mogu generalisati otežinjeni modularni grafovi data je u referenci [40]. Prvo se generiše binarna ER mreža (usmerena ili neusmerena) korišćenjem već opisanih prav-

ila. Zatim se čvorovi u ovako generisanoj mreži podele u G_o grupa. Težina linka se odredjuje iz intervala $\{a, a+1, \ldots, b\}$ ukoliko dva čvora pripadaju istoj grupi, odnosno $\{c, c+1, \ldots, d\}$ ukoliko čvorovi pripadaju različitim grupama. Prirodni brojevi a, b, c i d su odabrani tako da važi $c < d < a < b$. Jačina spregnutosti čvorova u mreži zavisi od razlike $a-d$, što je ova razlika veća to je modularna struktura mreže, koja je posledica različite težine linkova, izraženija.
Primena oMMV na usmerene i neusmerene otežinjene mreže generisane opisanim modelom za različite vrednosti parametra p i različite vrednosti brojeva a, b, c i d pokazuje da kao i u slučaju binarnih grafova, estimaciono-maksimizacioni algoritam brže konvergira za mreže sa većom gustinom i težinom linkova. Preciznost metoda zavisi od veličine mreže i odnosa težina linkova unutar i izmedju modula. Što je veća relativna težina linkova koji povezuju čvorove u istom podgrafu u odnosu na jačine linkova izmedju čvorova u različitim grupama, to je metod precizniji u pronalaženju ovih grupa čak i u gusto povezanim grafovima. Metod je testiran na grafovima sa velikim brojem linkova $p \geq 0.5$ i kao rezultat sledi da vrednost parametra p ne utiče na preciznost metoda. Rezultati dobijeni primenom oMMV na dve usmerene mreže veličine $N = 300$ sa istim brojem modula i istom raspodelom težina a različitim gustinom linkova ($p = 0.5$ i $p = 1$) pokazuju da je metod grupisao čvorove obe mreže u originalne grupe sa stoprocentnom tačnošću.
Na slici 3.6 je prikazana podela čvorova nadjena oMMV za $G = 3$ i $G = 5$ gusto povezanog grafa ($p = 0.5$) veličine $N = 300$ u kome originalno postoje $G_o = 5$ modula. Težine linkova koji povezuju čvorove unutar istog podgrafa birane su unifromno iz skupa $\{12, 13, \ldots, 18\}$ dok su težine linkova kojima su povezani moduli izabrane iz skupa $\{2, 3, \ldots, 8\}$. Kao i u slučaju binarnih grafova nadjeni podgrafovi za $G < G_o$ sadrže čvorove jedne ili više originalnih grupa, 3.6 (levo). Slika 3.6 (desno) pokazuje da su svi čvorovi svrstani u svoje originalne grupe. Ovde treba istaći da za razliku od modularnih binarnih mreža u ER otežinjenim grafovima nemamo centralne čvorove unutar modula, odnosno u srednjem svi čvorovi imaju iste iste *jačine* u odnosu na ulazeće i odlazeće linkove. Zbog velike gustine linkova unutar modula čvorovi imaju vrlo slične susede u odnosu na najjače odlazeće linkove, što je i jedan od razloga visoke preciznosti algoritma. Velika razlika izmedju težina linkova unutar i izmedju modula je garant da se prvi susedi odredjenog čvora nalaze u istom podgrafu što takodje utiče na samu efikasnost oMMV.
Slični rezultati su nadjeni i za neusmerene mreže, slika 3.7. Prikazana mreža je veličine $N = 300$ i gustine linkova $p = 0.7$, se sastoji od $G_o = 3$ modula koji su homogenije strukture u odnosu na jačine čvorova l_i u odnosu na slučaj prikazane usmerene mreže, slika 3.6. Nadjeni podgrafovi za $G = 3$ se podudaraju sa podelom čvorova u mreži koja sledi iz modela, slika 3.7.

Glava 4

Modularna struktura realnih otežinjenih mreža

Teorija kompleksnih mreža našla je primenu u različitim oblastima nauke, na primer biologiji, sociologiji, tehnologiji i t.d. Iako se sistemi koje izučavaju ove grane nauke drastično razlikuju, svi se ipak mogu svrstati u grupu kompleksnih sistema koji se mogu reprezentovati kompleksnim mrežama. Osobine ovih kompleksnih mreža se mogu direktno dovesti u vezu sa dinamikom i funkcijom samih sistema, što teoriju mreža čini moćnim alatom.
Većina realnih sistema se reprezentuje otežinjenim mrežama. Čvorovi u mreži mogu predstavljati različite subjekte (gene, filmove, postove, korisnike odnosno ljude) koje medjusobno interaguju. Težine linkova mogu biti različite prirode i zavise od posmatranog sistema. U biološkim mrežama težine linkova će predstavljati korelacije genskih ekspresija [45, 46], dok u mreži filmova *jačinu linka* predstavlja broj korisnika koji su ostavili komentar na oba filma [47, 18]. Nekada struktura sistema zahteva postojanje dva tipa čvorova, postovi/komentari i korisnici, gde je dozvoljeno vezivanje samo izmedju čvorova različitog tipa [18, 19].
Struktura i funkcija reprezentovanog sistema tesno je povezana sa topologijom mreže [1]. Pokazano je da mnoge realne mreže imaju nehomogenu strukturu kako na lokalnom tako i na mezoskopskom nivou [1, 46, 18]. Nalaženje otežinjenih podgrafova u mrežama je iz tih razloga jedan od važnih problema na koji se mora obratiti pažnja da bi se bolje razumela struktura a samim tim i funkcija dinamičkih sistema.
U ovoj glavi biće prikazana primena oMMV za nalaženje modula odnosno zajednica u biološkim i socijalno-tehnološkim sistemima. Iako konceptualno različiti ovi sistemi mogu biti reprezentovani i analizirani istom metodom, što samo svedoči o njegovom univerzalnom karakteru.

4.1 Mreža genskih ekspresija pivskog kvasca

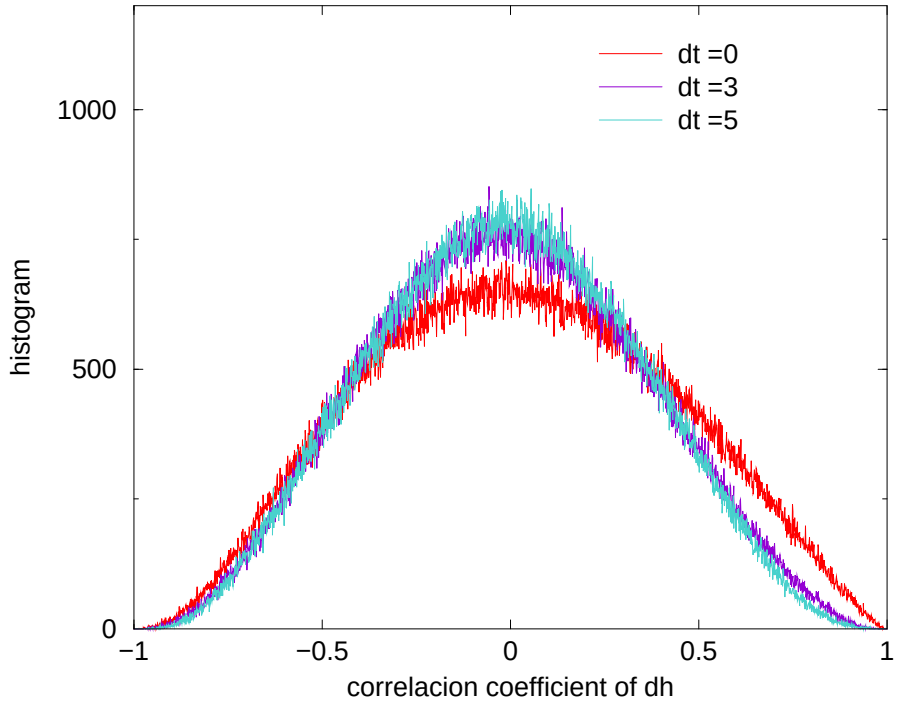
U radu [46] razmatrani su empirijski podaci za vremenske ekspresije čitavog genoma *pivskog kvasca*, *Saccharomyces cerevisiae*, merene u referenci [48] u 17 ekvidistantnih vremenskih trenutaka tokom dva puna ćelijska ciklusa. Statistička analiza podataka [45, 49], pokazuje njihovu invarijatnost u odnosu na skalu u rangiranju genskih ekspresija kao i odgovarajućim distribucijama, koje su karakteristične za samoorganizovane dinamičke sisteme. Na primer, rangirana srednja vrednost ekspresija gena tokom ćelijskog ciklusa ima oblik stepenog zakona (Zipov zakon) što ukazuje na različit značaj gena tokom ćelijskog ciklusa. Činjenica da geni imaju različit doprinos kolektivnom ponašanju čitave mreže je iskorišćena za odabir podskupa podataka od kojih će biti rekonstruisana mreža.
Analizirane su *diferencijalne ekspresije* gena definisane kao

$$\Delta X_i(t) = h_i(t) - h_i(t-1) \ , \tag{4.1}$$

za svaki od $N = 6406$ gena u bazi. Distribucija rangiranja za sve merene vrednosti (Nx17) kao za pojedinačne vremenske trenutke ukazuje na postojanje korelacija koje nisu slučajne izmedju gena čije su ekspresije veće od minimuma datog sa 2xh_0 (h_0 - srednja vrednost ekspresije za čitav sistem i za sve vremenske trenutke) [46]. Pod predpostavkom da imaju glavni uticaj na kolektivno ponašanje čitave mreže odabrani su oni geni čija srednja ekspresija zadovoljava uslov $< h_i >> 2h_0$. Na ovaj način dobijen je podskup od $N_s = 1216$ gena koji se mogu podeliti u dve grupe. Prvu grupu čine geni koji su aktivni tokom čitavog ćelijskog ciklusa ($N_1 = 612$) dok su preostali geni, njih $N_2 = 604$, aktivni tokom jedne ili eventualno dve od četiri faze ciklusa (G1, S, G2, M). Kako je prva grupa gena stalno aktivna njihove medjusobne korelacije će uvek biti nenulte, čineći grupu homogenom, t.j. bez nehomogene strukture. Druga grupa je iz tih razloga mnogo interesantnija, gledano sa aspekta postojanja modula, zbog čega će biti analizirana struktura njihove mreže.
Za svaki par gena može se izračunati korelacioni koeficijent izmedju diferencijalnih ekspresija $\Delta X_i(t)$ koji je dat formulom

$$C_{ij}(t-t') = \frac{\sum_t (\Delta X_i(t) - \langle \Delta X_i \rangle)(\Delta X_j(t-t') - \langle \Delta X_j \rangle)}{\sigma_i \sigma_j} \ , \tag{4.2}$$

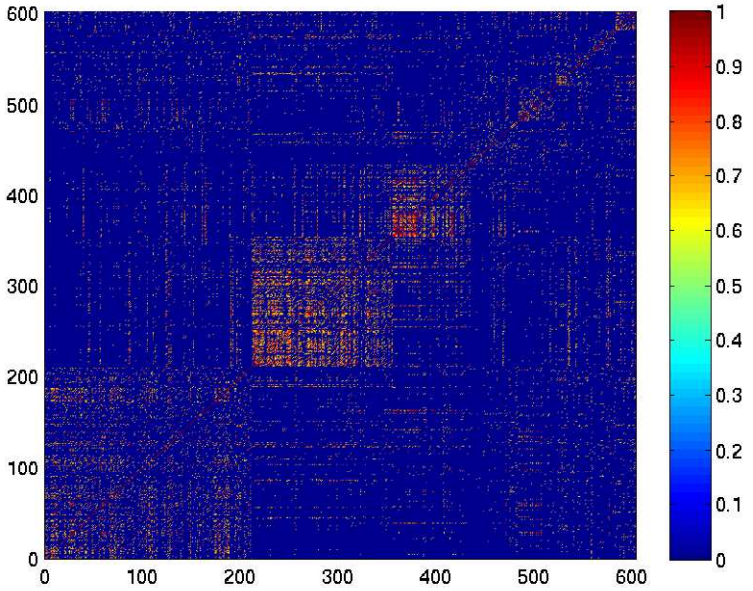
gde σ_i i σ_j predstavljaju standardne devijacije odgovarajućih vremenskih serija za gene i i j. Distribucija korelacionih koeficijenata za izabrani set od N_s gena je prikazana na slici 4.1, gde $dt = t - t^{'}$ označava vremenski pomak izmedju diferencijalnih ekspresija dva gena. Kao što se vidi na slici 4.1 najveće odstupanje od normalne distribucije imaju korelacioni koeficijenti za $\delta t = 0$, na osnovu kojih ćemo rekonstruistati odgovarajuću mrežu. Relevantne korelacije su one koje se nalaze u repu distribucije i da bi se odvojile od slučajnih, koje se nalaze u okolini nule, nophodno je zadati minimalnu vrednost korelacionog koeficijenta, W_0, i razmatrati



Slika 4.1: (levo) Distribucija korelacionih koeficijenata C_{ij} za $N = 1216$ gena za koekspresije ($dt = 0$) i dve koekspresije sa vremenskim zakasnjenjem.

samo korelacije čija je absolutna vrednost iznad ovog minimuma, $|C_{ij}| > W_0$. Kako ne postoje biološki argumenti za izbora korelacionog minimuma W_0, biće korišćen kriterijum koji proizilazi iz teorije kompleksnih mreža. Da bi se primenio metod oMMV neophodno je oderditi minimum W_0 tako da svi geni pripadaju jedinstvenoj povezanoj komponenti mreže, odnosno da izmedju svaka dva gena postoji *putanja* uticaja. U referenci [45] napravljena je sistemastka analiza formiranja jedinstvene povezane komponentne mreže u zavisnosti od W_0. Za odabrani set od N_s gena nadjeno je da je mreža povezana za W_0.6.

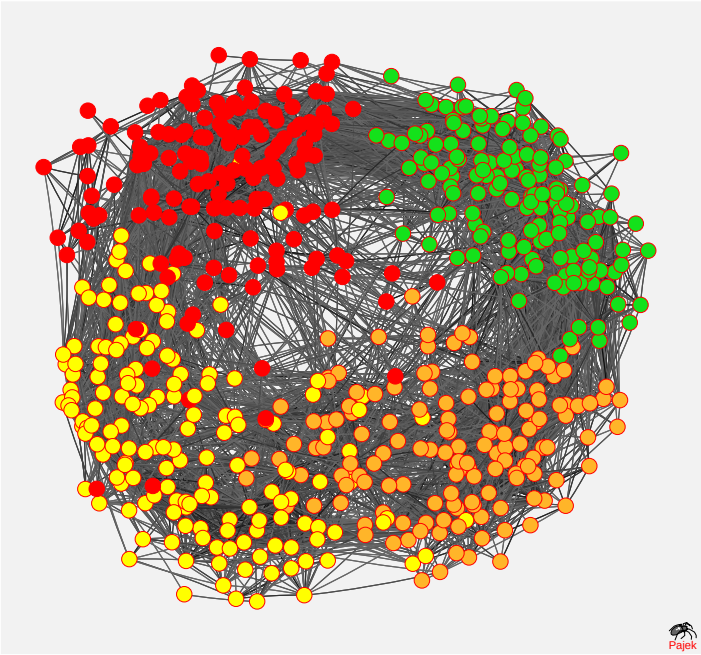
Mreža bazirana na korelacionoj matrici C sa elementima $C_{ij} > W_0$ je povezana, medjutim i pored uzimanja samo relevantnih korelacija broj linkova je i dalje velik. Veliki broj linkova, od kojih neki sa velikom težinom mogu biti lažno pozitivni, može uticati na tačnost dobijenih rezultata za strukturu podgrafova u mreži. Iz tih razloga neophodno je filtrirati korelacionu matricu korišćenjem metoda *metakorelacija* [50, 51], da bi se redukovao broj linkova. Metod *metakorelacija* se bazira na ideji da ukoliko su dva gena zaista korelisana, njihove korelacije sa ostalim genima u mreži će biti slične, t.j. korelacija izmedju njihovih korelacionih koeficijenata sa



Slika 4.2: Filrirana korelaciona matrica na kojoj se mže videti više modula gena ćelijskog siklusa ($N_{cc} = 603$).

ostalim genima u podskupu će biti bliska verdnosti 1. Da bi se primenio metod filtriranja neophodno je prvo transformisati interval korelacija C_{ij} sa $[-1, 1]$ na $[0, 1]$ korišcenjem formule $C'_{ij} = \frac{C_{ij}+1}{2}$. Na ovaj način se -1 korelacije slikaju u 0 dok korelacije 1 ostaju jednake 1, random korelacije su sada skoncentrisane oko vrednosti 0.5. Element filtrirane korelacione matrice C^M_{ij} je jednak proizvodu elementa C_{ij} iz stare matrice i odgovarajućeg elementa metakorelacione matrice M_{ij}. Element M_{ij} predstavlja korelacioni koeficijent kolona (redova) i i j matrice C a računa se na sledeći način: iz kolona C_{ik} i C_{jk} se izbace elementi C_{ii} i C_{jj} i kolone se urede tako da se elemnti C_{ij} odnosno C_{ji} nalaze na prvom mestu, $\{C_{ij}, C_{i1}, C_{i2}, ...\}$ i $\{C_{ji}, C_{j1}, C_{j2}, ...\}$. Ovako dobijeni vektori imaju $N_s - 1$ element. Zatim se izračuna Pirsonov korelacioni koeficijent M_{ij} ovako dobijenih vektora korišćenjem formule 4.2. Ukoliko su dva gena zaista korelisana, onda njihove korelacije sa ostalim genima prate isti obrazac, što za posledicu ima da je vrednost koeficijenta M_{ij} bliska 1. Vrednost elementa C^M_{ij} će u tom slučaju biti relativno veća od vrednosti ostalih elemenata u originalnoj matrici C_{ij}. U slučaju slučajnih korelacija dva gena $M_{ij} \simeq 0$, što za posledicu ima da je $C^M_{ij} \simeq 0$. Metodom filtriranja slučajne korelacije su pomorene ka nuli i nalaze se ispod korelacionog minimuma. Očekivano je da posle korišćenja metoda filtriranja korelacije izmedju sličnih gena, gena u is-

tom podgrafu, budu više izražene u odnosu na korelacije izmedju gena u različitim grupama. Ukoliko se sada sve korelacije $|C^M| < W_0 = 0.6$ odbace, dobija se filtrirana korelaciona matrica kao na slici 4.2. Sa slike se jasno vidi da mreži gena ćelijskog ciklusa veličine $N_2 = 604$ ima modularnu strukturu iako neki od modula nisu u potpunosti homogeni. Nehomogenost modula se može objasniti postojanjem gena sa visokom ekspresijom, što za posledicu ima jaku korelisanost tog gena sa genima kako u sopstvenom modulu tako i u ostalim podgrafovima. Ove korelacije su posledica postojanja hijerarhijske strukture čvorova u modulima (postojanje centralnih čvorova) i ne mogu se eliminisati primenom standardnih metoda filtriranja [51]. Ipak kao što je pokazano u referenci [51], postojanje ovih korelacija ne utiče na rezultate koji se odnose na veličinu i pripadnost čvorova odredjenom modulu u slučaju modeliranih modularnih mreža. Za očekivati je da isto važi i za korelacije gena.



Slika 4.3: Grupe gena u mreži pivskog kvasca nadjene oMMV algoritmom za $G = 4$.

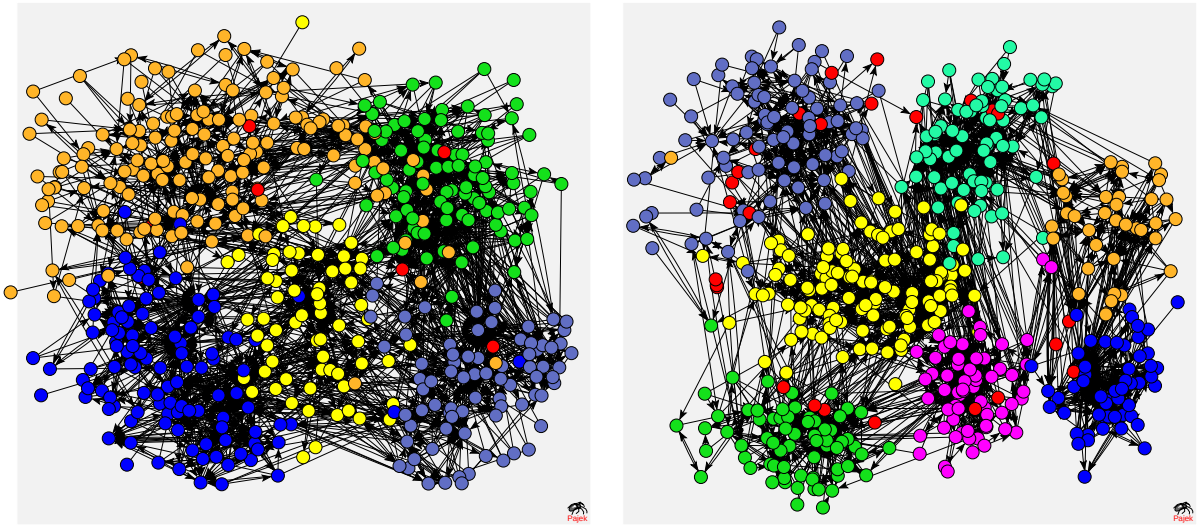
Primenom oMMV metoda na mrežu generisanu iz filtrirane korelacione matrice nadjenje su $G = 4$ grupe gena približno jednake veličine, 4.3. Pri reprezentovanju korelacione matrice odbačeni su svi linkovi sa težinom manjom od $W_0 = 0.6$ što za posledicu ima da dobijena mreža nije poptpuni graf (srednji broj linkova po čvoru je $L = 32$). Broj grupa u mreži je odredjen korišćenjem metoda spektralne analize [46]. Geni u različitim modulima su grupisani prema njihovom položaju

u ćeliji: mitohondrija (crvena), periferija ćelije (zelena), jedro (žuta) i citoplasma (narandžasta). Slična podela je dobijena i primenom metoda spektralne analize [46].

4.2 IMDB

Reprezentovanje socijalnih interakcija grafovima [52, 18], predstavlja klasičan primer primene teorije kompleksnih mreža. Jedna od karakterističnih osobina klasičnih socijalnih kao i internet zajednica je grupisanje ljudi koje se u grafovima manifestuje kao podgraf ili zajednica. Grupisanje ljudi u običnim zajednicama je bazirano na različitim socijalnim kategorijama, kao što su prijateljstvo, profesionalni aspekt, škola, geografsko poreklo, t.j. postoji neka spona izmedju ljudi koji se nalaze u odredjenoj grupi. Za razliku od klasičnih socijalnih kontakata internet zajednice su karakterisane *odsustvom ličnog kontakta*, velikim arhivama sa neograničenim pristupom, brzim i simultanim interakcijama i nepostojanjem ograničenja uslovljenim geografskim položajem. Svaka od ovih karakteristika može uticati na osobine ponašanja socijalnih grupa i društva uopste. Nagli razvoj veb portala kao što su Fejsbuk, Majspejs ili različiti portali koji se odnose na odredjene sadržaje (filmska baza -IMDb) zahtevaju a ujedno zbog svojih arhiva i omogućavaju izučavanje ponašanja ljudi primenom teorije mreža.
U radu [47, 18] analizirane su podaci filmska baze - IMDb [41]. IMDb je veb portal osmisljen sa idejom da obezbedi što potpunije informacije o odredjenom filmu i time korisniku omogući lakš izbor filma. Osim uobičajenih informacija o filmu, korisniku se nudi lista preporuka od deset filmova odabranih tako da budu u odredjenoj vezi sa posmatranim filmom, na primer isti žanr ili režiser odnosno glavni glumac. Pri izboru filma korisnik se može koristiti i komentarima koje su ostavljali drugi korisnici. Registrovani korisnici na IMDb portalu formiraju komunu sa svim karakteristikama socijalnih veb zajednica.
Analizirani podaci se odnose na filmove američke produkcije. Za svaki od filmova postoji sledeći skup informacija: jednistvena šifra filma (dodeljena od strane baze), naslov, žanr i podžanrovi, broj glasova, ocena i lista korisnika koji su ostavili komentar na film (identitet korisnika nije poznat samo njegov nadimak). Ovde treba istaći da u podacima postoje dva tipa subjekata koji interaguju: korisnici i objekti njihove interakcije (filmovi). Podaci ovakve strukture se standardno reprezentuju bipartitnim mrežama. Interakcije izmedju filma i korisnika predstavlja komentar ostavljen od strane korisnika koji se odnosi na film. Veličina podskupa filmova (N_M) i korisnika (N_U) pa samim tim i večina mreže ($N_M + N_U$) zavisi od načina filtriranja podataka. Kao filter parametar korišćen je minimalni broj glasova, V_{min}.
Bipartitna mreža je samo jedna od mogućih reprezentacija podataka pogodnih za analizu. Analizom različitih monopartitnih mreža koje se mogu konstruisati iz po-

dataka i iz bipartitne mreže moguće je dobiti relevantne informacije o ponašanju korisnika i o strukturi same baze. Mreža konstruisana direktno iz liste preporuka je **umserena mreža filmskih preporuka** (UMFP). Čvorovi u mreži, filmovi, su medjusobno povezani usmerenim linkovima koji idu od filma i ka filmovima koji se nalaze na njegovoj listi preporuka. Ove liste su iz praktičnih razloga ograničene na 10 filmova, što znači da iako se odredjeni film k nalazi na listi preporuka filma ℓ, zbog čega postoji link $\ell \to k$, film ℓ ne se ne mora se nužno nalaziti na listi preporuka filma k. Ovo ima za posledicu usmereni karakter linkova u mreži UMFP.



Slika 4.4: Primer modularne strukture u mrežama filmova: (levo) mreža dobijena direktno iz podataka *IMDB* i (desno) binarne projekcije bipatritne mreže. Vrednost parametra $V_{min} = 30000$.

Sličnu usmerenu mrežu moguće je konstruisati iz normalizovane projekcije bipartitne mreže na grupu filmova,(NPBMf). U projektovanoj mreži link izmedju dva filma i i j postoji ukoliko oni imaju zajedničkog korisnika k, t.j. dva linka ik i jk se projektuju na jedan link ij. Na ovaj način je veličina mreže redukovana na N_M. Jasno je da dva filma mogu imati više od jednog korisnika, što za posledicu ima postojanje višestrukih ili otežinjenih linkova medju njima. Jedan od relevantnih načina za ocenu težine linkova je razmatranje skalarnog proizvoda vektora $\vec{\mu_i}$ komentara korisnika na film i:

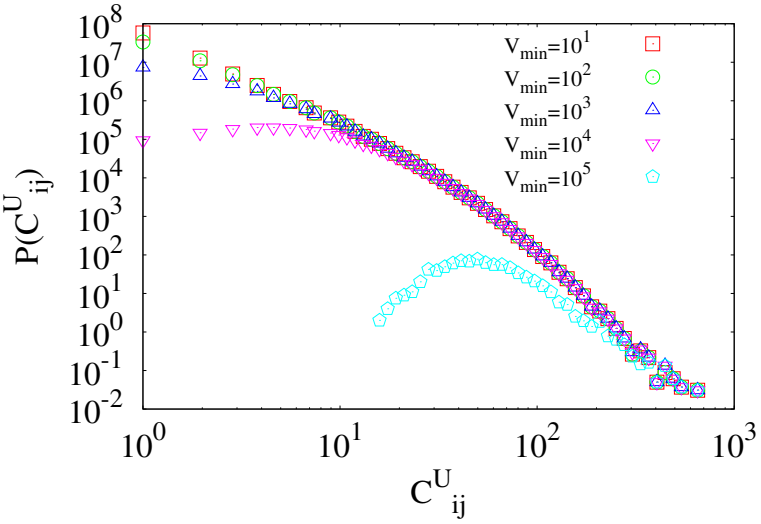
$$\vec{\mu_i} = (\ldots, 1, \ldots, 0, \ldots, 0, \ldots, 1, \ldots) \qquad .$$

j-ta komponenta vektora $\vec{\mu_i}$ je jednaka 1 ukoliko je korisnik j ostavio komentar na film i. Težina linka je data normalizovanim *skalarnim proizvodom* izmedju filmova

i i j

$$d_{ij} = \frac{\overrightarrow{\mu_i} \cdot \overrightarrow{\mu_j}}{|\mu_i||\mu_j|} \ . \tag{4.3}$$

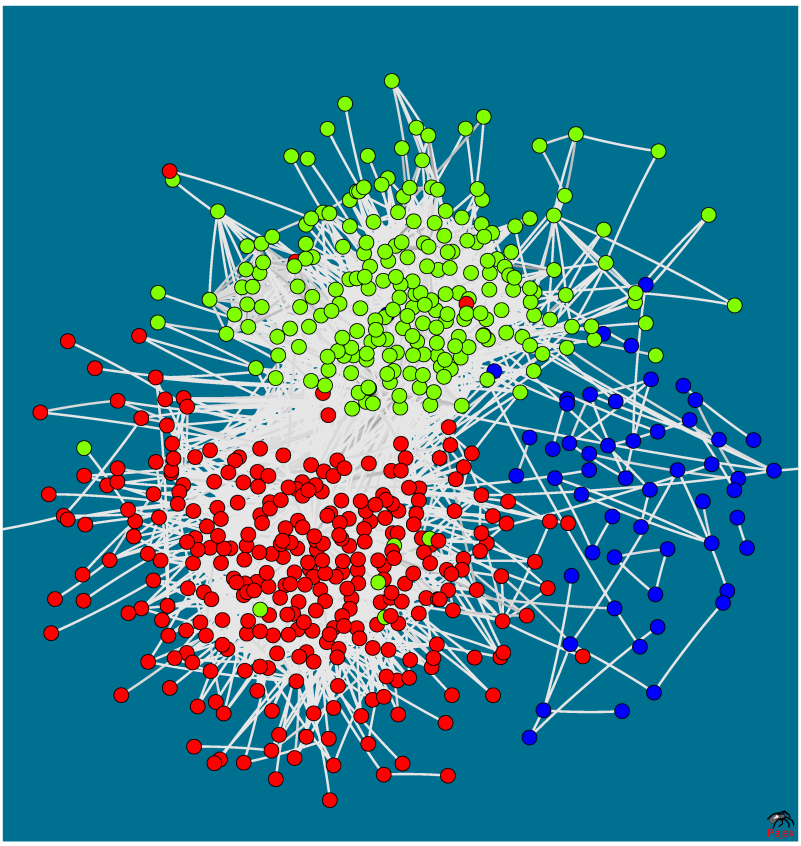
Ovako projektovana mreža je simetrična, medjutim ukoliko se broj filmova (linkova) ograniči na na primer 10 onih sa najvećim težinama po čvoru, mreža postaje usmerena. Dobijena otežinjena usmerena mreža može se pretvoriti u binarnu ukoliko zanemarimo težine linkova. Na ovaj način se dobija mreža slična mreži preporuka čija su struktura i osobine uslovljene ponašanjem korisnika.



Slika 4.5: Distribucija broja zajedničkih korisnika po paru filmova za različite vrednosti parametra V_{min}.

Osobine kao što su distribucija povezanosti i asortativnost mreža UMFP i NPBMf ukazuju na njihovu sličnu strukturu na lokalnom nivou [18]. Distribucija verovatnoće dolazećih linkova za velike stepenove čvora, $q_M \geq 10$, ima oblik stepenog zakona $P(q) \sim q^{-\tau_M}$ sa vrednošću eksponenta $\tau_M \simeq 2$ za obe mreže. Ovakav oblik distribucije povezanosti ukazuje na postojanje hierarhijske strukture filmova u mreži, odnosno da je u mreži prisutan mali broj filmova koji se nalaze na listama preporuka velikog broja filmova. Korelisanost izmedju čvorova u mreži se može opisati srednjim stepenom čvorova, $< q >_{ps}$, prvih suseda čvora sa stepenom q. Funkcija $< q >_{ps}$ raste sa porastom q što u slučaju UMFP mreže znači da ukoliko je film često preporučivan velika je verovatnoća da će se na njegovoj listi naći često preporučivani filmovi. Za mreže kod kojih se čvorovi vezuju za čvorove sa sličnim stepenom se kaže da su asortativne. Kod NPBMf je asortativnost direktna posledica ponašanja

vrlo aktivnih korisnika.
Iako ispoljavalju slične lokalne osobine, grupisanje filmova u ovim mrežama je drugačije. Na slici 4.4 prikazan mreže UMFP i NPMB za filmove sa minimalno $V_{min} = 30000$ glasova veličine $N_M = 518$ čvorova. Spektralna analiza pokazuje različit broj grupa filmova u mrežama, $G = 5$ za UMFP odnosno $G = 7$ za NPMB. Sastav grupa nadjen je primenom MMV za usmerene mreže, topološki moduli označeni su različitim bojama, 4.4.



Slika 4.6: Otežinjena mreža filmova sa brojem glasova izmedju 1000 i 2000. Na slici prikazan samo deo mreže veličine 596 čvorova. Mreža je sečena prema težinama linkova $C_{ij}^M > 5$. Boje predstavljaju različite grupe čvorova nadjenje oMMV algortimom: čvorovi u crvenoj grupi su horor filmovi, u zelenoj drame a u plavoj grupi su čvorovi koji se svrstavaju u ljubavni žanr

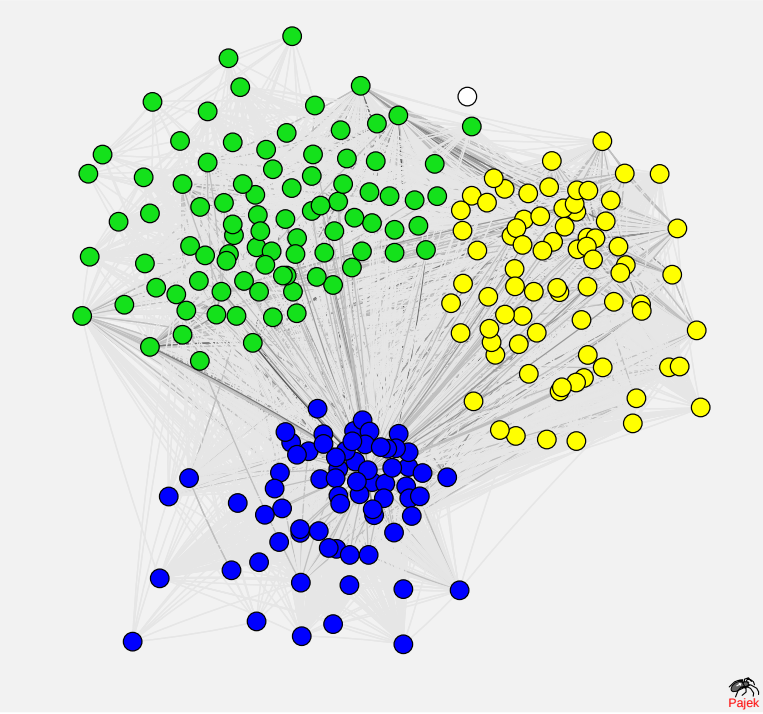
Otežinjene projekcije bipartitne mreže na grupu filmova (OF) odnosno korisnika (OK) su još jedan način reprezentovanja podataka sa IMDb-a. Ponašanje korisnika, posebno njihovo grupisanje, direktno utiče na modularnu strukturu oteženjenih pro-

jekcija. U otežinjenoj mreži korisnika, dva korisnika su povezana ukoliko su ostavili komentar na isti film. Višestruki linkovi izmedju istog para korisnika se javljaju ukoliko su korisnici ostavili komentar na više istih filmova. U ovom slučaju je veličina mreže redukovana na N_U korisnika. Slično u otežinjenoj mreži filmova, težina linkova izmedju dva korsnika je proporcionalna broju zajedničkih korisnika. Veličina OF mreže je jednaka N_M. Raspodela verovatnoće zajedničkih korisnika, C_{ij}^U po paru filmova za različite vrednosti parametra V_{min} prikazana je na slici 4.5. Distribucija ima oblik stepenog zakona za velike vrednosti C_{ij} nezavisno od veličine mreže. Jedina uočljiva razlika je za jako popularne filmove sa više od $V_{min} = 100000$ glasova gde funkcija $P(C_{ij}^U)$ ima jako velik nagib.

Prilikom projekcije bipartitne mreže na monopartitni graf priroda vezivanja čvorova u grafu se drastično menja. Tipična lokalna struktura zvezde u bipartitnom grafu, na primer 5 filmova povezanih za jednog korisnika, se u mreži filmova projektuje na potpuno povezani podgraf od 5 filmova. Iz tih razloga projektovane mreže su jako guste, sa velikim srednjim brojem linkova po čvoru.

Iskoristili smo metod za otežinjene mreže za identifikovanje grupa u mreži filmova srednje popularnosti, sa brojem glasova u intervalu $[1000, 2000]$. Na ovaj način smo redukovali veličinu mreže na $N_M = 1510$ čvorova. Spektralna analiza odgovarajućeg normalizovanog Laplasijana pokazuje da u mreži postoje tri uslovno rečeno odvojene grupe čvorova [18]. Na slici 4.6 prikazana je podela mreže nadjena oMMV metodom, čvorovi sa istom bojom pripadaju istom podgrafu. Velika gustina linkova u analiziranoj otežinjenoj mreži onemogućava da se ona prikaže cela. Radi ilustracije rezultata metoda originalnu mrežu od $N_M = 1510$ čvorova smo odeskli odbacivši sve linkove težine manje od $C_{ij}^U < 6$. Zbog postojanja slabo povezanih čvorova (svi linkovi imaju težinu $C_{ij} <= 5$) veličina mreže je takodje redukovana na 595 čvora.

Podaci o žanrovima i podžanrovima filmova omogućavaju nam da kvalitativno opišemo grupe nadjene primenom oMMV. Analizirani filmovi se mogu podvesti pod tri glavna žanra: horor, romansa i komedija. Grupe nadjene oMMV kao i struktura same mreže ukazuju na podelu korisnika na dve grupe prema njihovom ponašanju. Prvu grupu korisnika čine ljudi koji komentarišu (gledaju) filmove nezavisno od njihovog žanra. Njihovo ponašanje doprinosi velikoj povezanosti u mreži u smislu da postoji relativno mali broj čvorova koji medjusobno nisu povezani bar jednim linkom. U drugu grupu spadaju korisnici koji uglavnom gledaju i ostavljaju komentare na filmove odredjenog žanra. Upravo njihova aktivnost ima za posledicu veliku povezanost filmova iste tematike. Čvorovi iste boje na slici 4.6 predstavljaju filmove istog žanra: horor filmovi su reprezentovani čvorovima crvene boje, zelene boje su drame dok su u plavoj grupi romanse. Slična podela filmova po grupama nadjena je i metodom spektralne analize [18].



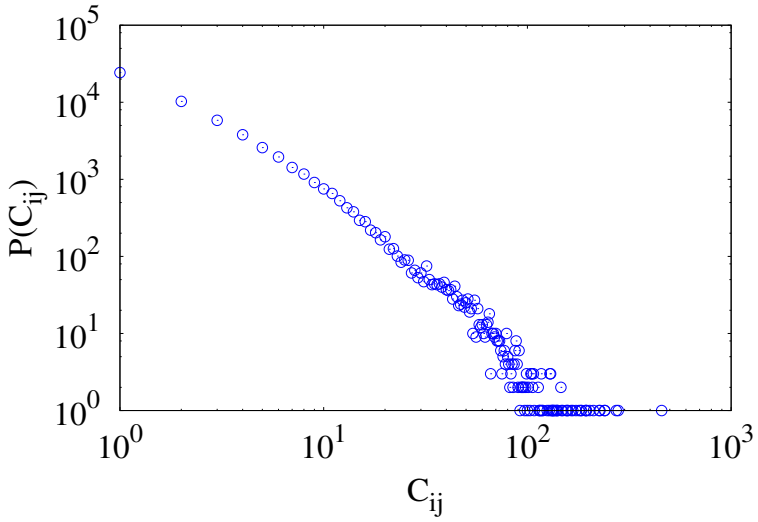
Slika 4.7: Grupe postova pronadjene u otežinjenoj mreži normalnih postova bloga B92 korišćenjem oMMV algoritma za $G = 3$. Identifikovane zajednice čine postovi različite tematike.

4.3 Osobine mreža socijalne zajednice bloga B92

Još jedan primer socijalnih internet zajednica predstavljaju blogovi i forumi. Prvenstevno osmisljeni kao *mesta* gde ljudi mogu da iznose i razmenjuju svoja mišljenja, blogovi su postali mesta "okupljanja" ljudi u sajber prostoru. Sve češće smo svedoci organizovanja ljudi na blogu koja mogu imati uticaj na naš svakodnevni život, na primer gradjanska inicijativa "Majka Hrabrost" čiji su počeci vezani za blog B92 [42]. Za razliku od baze filmova gde su korisnici ograničeni na ostavljanje komentara na odredjeni film, na blogovima su upravo korisnici ti koji biraju teme o kojima će se diskutovati. Blog je mnogo dinamičniji od IMDb baze upravo zbog činjenice da se sva dešavanja u svetu i životima njegovih članova pretaču u postove i komentare. Diskusije izmedju ljudi oprečnih stavova su neretko nabijene emocijama i mogu eskalirati u sukobe. Članovi blog zajednica često nastupaju pod pseudonimima što utiče na njihovo ponašanje. Skrivenost identiteta ih čini smelijim i manje podložnim moralnim i etičkim normama koje prisutne u svakodnevnoj komunikaciji. Ovo ima za posledicu da su obrasci ponašanja ljudi u ovim zajednicama drugačiji od ponašanja u svakodnevnoj komunikaciji, "lice u lice". Upravo ova činjenica kao i rast popularnosti blog zajednica kao jednog od vidova komunikacije izmedju ljudi

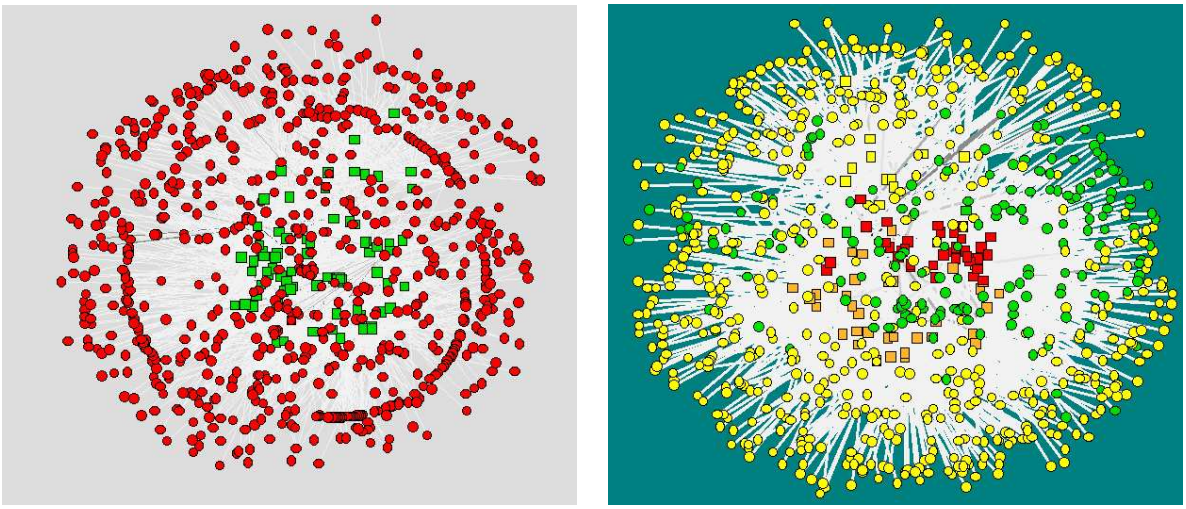
nameće potrebu za izučavanjem njegovih osobina. Metodi teorije mreža mogu pružiti neophodne informacije o obrascima ponašanja ljudi na blogu odnosno o načinu formiranja grupa korisnika.
Jedan od najpopularnijih i najdinamičnijih blogova u Srbiji je blog sajta B92 [42]. Ovaj blog je počeo sa radom krajem maja 2007 godine, od kada beleži neprekidni rast broja korisnika. Analizirani podaci su za period od skoro dve godine, od 27. maja 2007 do 01.03. 2009 godine, što odgovara grupi od $N_U = 4598$ korisnika koji su ukupno ostavili $N_P = 4784$ postova i $N_C = 406527$ komentara. U razvoju bloga B92 može se razlikovati više faza. U početku je blog bio zatvorenog tipa (do jula 2008), odnosno samo ljudi sa odgovarajućom pozivnicom su mogli da postanu članovi, od kojih su opet samo odabrani imali VIP status i autorsku opciju (mogućnost pisanja postova). Tokom vremena svi članovi su dobili autorsku opciju, ali podela na VIP i obične autore je prisutna i danas. Članstvo u VIP grupi nije fiksirano, odnosno moguće je dobiti ili ostati bez VIP opcije, što zavisi od uredničke politike bloga. Set analiziranih podataka se odnosi na postove koje su napisali korisnici koji su bar jednom imali VIP opciju, odnosno njihovi postovi su bar neko vreme bili istaknuti na glavnoj strani bloga. Od jula 2008 godine, blog je otvoren i svako može da se registruje i postane član.
Podatake sa bloga je takodje moguće reprezentovati bipartitnom mrežom u kojoj se mogu razlikovati dve grupe čvorova. Prvu grupu čine korisnici dok u drugu grupu spadaju postovi i komentari ostavljeni na postove. Za razliku od baze filmova gde korisnik samo ostavlja komentar na odredjeni film, u blogovima korisnik može biti autor posta(komentara) ali i može komentarisati post. Da bi se razlikovalo kada je korisnik autor a kada čita odredjeni post(komentar) neophodno je uvesti usmerenost linkova u bipartitnoj mreži na sledeći način: link ide od posta i ka korisniku koji ga čita k, $i \to k$ i ostavlja komentar l, sto se reprezentuje linkom $k \to l$. Na blogu b92 je moguće ostaviti i komentar na komentar. Ovo znači da je korisnik k pročitao odredjeni komentar l_1 ali i odgovarajući post i, situacija koja se reprezentuje linkovima $l_1 \to k$ i $i \to k$.

Za bipartitnu mrežu blogova moguće je odrediti raspodelu broja dolazećih , q_u^U, i odlazeíh, q_i^U, linkova za korisnike, i odlazećih za postove i komentare, q^B, (svaki post/komentar ima po jedan dolazeći link, t.j. jednog autora). Raspodela povezanosti za obe vrste stepenova koji se odnosi na korisnike imaju oblik stepenog zakona, $P(q) \sim q^{-\tau}$, za male i srednje vrednosti q, sa eksponentom $\tau \sim 1.5$ [19]. Distribucija verovatnoća za q^B ima takodje oblik stepene raspodele sa dve različite vrednosti eksponenata. Za male i srednje vrednosti q^B, $\tau^B \sim 2$, dok za velike vrednosti $q > 100$ eksponent τ je približno jednak 3.5. Ovakav oblik raspodele ukazuje na postojanje dve grupe postova, takozvanih normlnih koji imaju manje od 100 komentara i popularnih. Ponašanje korisnika vezano za ove dve grupe postova se drastično razlikuje, zbog čega se one moraju analizirati odvojeno.
Razmatraćemo dva tipa mreža koje su vrsta projekcije bipartitne mreže. Otežinjena



Slika 4.8: Distribucija broja komentara koje jedan korisnik ostavi na odredjeni post za grupu popularnih postova. Broj komentara koje jedan korisnik ostavi na neki post je jačina linka izmedju korisnika i posta u bipartitnoj mreži.

mreža korisnika na blogu dobija se na sličan način kao i mreža korisnika baze filmova. Dva korisnika su povezana linkom ukoliko su ostavili komentar na isti post (komentar). Višestrukost linkova se može pojaviti ukoliko dva korsnika ostave komentar na više različitih postova(komentara) ili ukoliko ostave više komentara na isti post (komentar). Mreža korisnika je dobijena iz bipartitne reprezentacije podskupa podataka koji se odnose na normalne postove, broj komentara manji od 100. Veličina projektovane mreže jednaka je $N_U = 3367$ korisnika koji su komentarisali normalne postove. Raspodela težina linkova u ovoj mreži ima oblik stepene raspodele sa eksponentom $\tau \simeq 2$. Iz spektralne analize normalizovanog Laplasijana koji odgovara ovoj mreži [19], sledi da su korisnici grupisani u 4 grupe oko postova različite tematike. Za dve manje populisane grupe korisnika je moguće odrediti temu na osnovu naslova i sadržine postova koje su komentarisali. Zbog veličine za druge dve grupe nije moguće primeniti isti metod, već je neophodno ići u dublju analizu strukture podgrafova korišćenjem teorije mreža. Radi ilustracije kako oMMV metod može primeniti na odredjivanje postova oko kojih se grupišu korisnici, odabrali smo jednu od ovih grupa od $N = 47$ korisnika. Zatim smo odredili podgrupu postova koje su komentarisali ovi korisnici $N_P = 236$ postova i od nje i odgovarajućih komentara, $N_C = 11110$, napravili bipartitnu mrežu. Projekcijom ove mreže na grupu postova dobija se otežinjena mreža, gde težina linka izmedju dva posta odgovara broju njihovih zajedničkih korisnika. Primenom oMMV na ovu mrežu za $G = 3$ nadjene su grupe postova sa različitom tematikom, slika 4.7.



Slika 4.9: Grupe čvorova nadjene oMMV algoritmom u otežinjenoj bipartitnoj mreži za $G = 2$ (levo) i $G = 4$ (desno).

Podaci koji se odnose na popularne postove su analizirani na drugačiji način od podataka za normalne postove. Analizirana je otežinjena neusmerena bipartitna mreža od $N_p = 1466$ postova i $N_U = 3613$ korisnika. Kao i u slučaju binarne bipartitne mreže i u ovoj mreži ne postoje linkovi izmedju čvorova unutar iste particije. Link izmedju posta i i korisnika k postoji ukoliko je korisnik ostavio komentar na odredjeni post ili je korisnik napisao post. Težina linka $i \to k$ je proporcionalna broju komentara koje je korisnik k ostavio na post i. Na slici 4.8 prikazana je raspodela težine linkova ove mreže stepenog oblika sa eksponentom $\tau \simeq 3$. Da bi smo redukovali veličinu mreže, radi primene oMMV, napravljena je prvo spektralna analiza mreže. Iz spektralne analize sledi da postoje 6 grupa postova i korisnika od kojih smo odabrali dve. Zatim je od njih napravljena otežinjena bipartitna mreža koja sadrži 87 postova i 744 korisnika.

Korišćenjem metoda dobija se nekoliko podela mreže za različite vrednosti parametra G. Za $G = 2$ kao i u slučaju binarnih bipartitnih mreža [34], metod nalazi podelu grafa na particije, t.j. na korisnike i postove, slika 4.9 (levo). Za veće vrednosti parametra G neke od nadjenih grupa su mešovite, t.j. ne postoji više jasna podela na korisnike i postove, slika 4.9 (desno). Rezultati za $G = 4$ pokazuju da je oMMV mnogo uspešniji u razdvajanju particije postova nego particije korisnika. Postovi su podeljeni u 4 grupe, od čega dve grupe sadrže samo postove (crvena i narandžasta) dok se u preostale dve (zelena i žuta) nalaze i postovi i korisnici. Narandzastu grupu čine postovi koji se odnose na inicijativu "Majka hrabrost" i stanja u srpskim porodilišitima, dok se postovi obojeni crvenom bojom nemaju neku odredjenu tematiku ali su pisani od strane četvoro autora. Korisnici koji su uglavnom komen-

tarisali ove postove se nalaze u žutoj i zelenoj grupi. Zelenu grupu pored njih čine i postovi koji se bave političkom scenom u Crnoj Gori i Srbiji kao i korisnici koji su ih komentarisali. Postovi stavljeni u žutu grupu nemaju nikakvih zajedničkih karakteristika i mogu se smatrati greškom metoda. Ovde treba istaći da se povećanjem vrednosti parametra G ne dobija bolja podela korisnika po grupama.
Podela mreže na 4 grupe, koja se delimično slaže sa rezultatima oMMV, nadjena je i spektralnom analizom normalizovanog Laplasijana [19]. Za razliku od oMMV grupe nadjene spektralnom metodom sadrže i odgovarajuće korisnike. Kako su grupe nadjene spektralnom metodom mešovite, podelu na korisnike i postove unutar grupa je moguće naći primenom oMMV.

Glava 5

Zaključak

U ovom radu predstavljen je generalizovani metod baziran na maksimizaciji funkcije verodostojnosti za nalaženje podgrafova u usmerenim i neusmerenim otežinjenim mrežama. Pokazano je da je metod za nalaženje modula u binarnim grafovima samo specijalan slučaj generalizovanog metoda u slučaju kada se težine linkova zanemare. Predstavljena je implementacija algoritma sa osvrtom na detalje koji se odnose na odabir inicijalne podele mreže i broj neophodnih ponavljanja algoritma kako bi se pronašao odgovarajući minimum.
Primena metoda na kompjuterski generisane binarne i otežinjene mreže sa dobro definisanom i poznatom modularnom strukturom poslužila je kao test efikasnosti i preciznosti metoda. Binarne mreže gnerisane su iz modela modularnih mreža čija se struktura i broj modula kontroliše parametrima M, α i P_0. Efikasnost metoda zavisi od veličine i broja podgrafova u mreži. Za $G = G_o$ metod uspešno identfikuje grupe čvorova u binarnim usmrenim mrežama koje sadrže $N = 1000$ čvorova ukoliko je broj podgrupa manji od 10. Pokazano je da je postepenim povećavanjem parametra G moguće identifikovati hijerarhijsku modularnu strukturu mreže. Metod je efikasan u nalaženju čvorova konektore u mreži, medjutim njegova primena na za identifikaciju grana u drvetu drveta kao dve različite particije čvorova nije moguća. Primena metoda na Erdoš Renji generalisanom modelu otežinjenih grafova pokazuje da je metod uspešan u nalaženju modula u vrlo gustim mrežama i na potpuno povezanim grafovima. Brzina konvergencije algoritma zavisi od gustine i težine linkova u mreži, dok preciznost metoda zavisi od odnosa težina linkova u i izmedju podgrafova. Metod je podjednako uspešan u identifikovanju zajednica u usmerenim i neusmerenim mrežama. Kao i u slučaju binarnih grafova, nalaženjem grupa čvorova u mreži za različite vrednosti parametra G identifikuje se hijerarhijska struktura podgrafova.
Maksimalna vrednost funkcije verodostojnosti raste sa povećanjem parametra G zbog čega nije moguće iz metoda odrediti idelanu podelu mreže. Za ocenu broja podgrafova u mreži je neophodno koristiti neki drugi metod, na primer spektralnu anal-

izu ili vrednost modularnosti. Kao posledicu kompleksne strukture mreža imamo da funkcija maksimalne verodostojnosti može imati veliki broj maksimuma, zbog čega je neophodan relativno veliki broj ponavljanja. Nadjeni maksimum i odgovarajući parametri modela predstavlja najbolju ocenu podele mreže na zajednice.
U četvrtoj glavi prikazana je primena metoda na nalaženje modularne strukture u bioloških i socijalnih mreža generisanih iz podataka. Iz korelacione matrica genskih ekspresija pivskog kvasca moguće je generisati otežinjenu neusmrenu mrežu. Analiza mreže podskupa gena ćelijksog ciklusa pokazuje da su ovi geni grupisani u četiri modula. Identifikovane grupe odgovaraju genima koji su aktivni u različitim fazama ćelijskog ciklusa.
Primena metoda na socijalne veb mreže nam može pružiti neophodnu informaciju o ponašanju članova oderedjene veb zajednice. Podaci sa IMDb filmske baze, reprezentovani bipartitnom mrežom i odgovarajućim monopartitnim projekcijama, ispoljavaju ispoljavaju modularnu strukturu. Poredjenje dve binarne usmerene mreže filmova dobijene iz relanih podataka i porjekcijom binarne mreže ukazuje da na različitu prirodu formiranja ovih mreža. Ponašanje korisnika indukuje različitu modularnu strukturu ovih mreža nadjenu korišćenjem metoda maksimalne verodostojnosti. Grupe filmova najdene u otežinjenoj mreži su tesno povezane sa njihovim žanrovima, na osnovu čega se može zaključiti da većina korisnika IMDb komentariše i gleda filmove odredjenog žanra.
Ponašanje korisnika na blogovima utiče na njegovu strukturu što se ogleda u raličitim karakteristikama bipartitne mreže. Analiza podmreže grupe korisnika koji su u vezi sa normalnim postovima (manje od 100 komentara), pokazuje da se za razliku od filmova korisnici ne grupišu oko postova slične tematike već da njihovo ponašanje zavisi mnogo više od toga ko je autor posta.
Na osnovu svega navedenog, može se zaključiti da se metod može uspešno primeniti za nalaženje modularne strukture u cikličnim i gustim mrež kojima su opisani sistemi različite prirode i porekla.

Dodatak A

Programski kod

Programski kod za numeričko nalaženje zajednica u otežinjenim i binarnim mrežama baziran na metodu maksimalne verodostojnosti razvijen je u jeziku C++, uz poštovanje ANSI standardizacije. Tipično vreme izvršavanja koda za retku mrežu od $N = 1000$ čvorova sa 10% prevezanih linkova, $G_o = 4$ modula i gusitne povezanosti $M = 3$ je reda veličine par minuta na Intel Centrino 2.00 GHz platformi je reda veličine par minuta. Ulazni parametri su veličina mreže, n, broj modula, G, kontrolni parametar *control* i *seed* kojim se inicijalizuje SPRNG, generator slučajnih brojeva kao i matrica težina W. Matrica trežina je zapisana u ulaznom fajlu kao niz linija gde svaka sadrži podatke o jednom linku u obliku ijW_{ij} (i *izvor linka*, j *meta linka* i W_{ij} ˇtežina). Izlazni podaci su razvrstani u pet fajlova: *likeC* sadrži procenjenu vrednost maksimuma verodostojnosti, *likelihood* sadrži vrednosti funkcije verodostojnosti za svaki od iteracionih koraka, *p*, *teta*, *q* fajlovi sadrže konačne vrednosti parametara modela i vrednosti verovatnoća q_i, dok fajl *cluster.clu* sadrži podelu mreže i u formi je *pajek* fajla.

```
#include <math.h>
#include <sprng.h>
#include <iostream>
#include <fstream>

using namespace std;

int streamnum, nstreams, *stream;

long seed;

double **W;
double **q;
double *pi;
```

```
double **teta;
double *pio;
double **tetao;
double *epom;
short int *epomc;
short int *gindex;
double *strenght;
short int G;

int v1, v2;
double control;

double PI,pom,L,x,x1,x4,x2,x3,z,y,eps,ksi2,ksi1;
short int a,nmin,sum,n;
short int p,stop2,stop1,stop,i,j,r,s,st;

char sused[500];

main (int argc,char **argv)
{

n=atol(argv[1]);
G=atol(argv[2]);
seed=atol(argv[4]);
control=atof(argv[3]);

W=new double*[n+1];
for(i=1;i<=n;i++) W[i]=new double[n+1];

q=new double*[G+1];
for(i=1;i<=n;i++) q[i]=new double[n+1];

pi=new double[G+1];

teta=new double*[G+1];
for(i=1;i<=n;i++) teta[i]=new double[n+1];

pio=new double[G+1];

tetao=new double*[G+1];
for(i=1;i<=n;i++) tetao[i]=new double[n+1];

epom=new double[G+1];
epomc=new short int[G+1];
```

```
gindex=new short int[n+1];

strenght=new double[n+1];



ofstream f,f1,f2,f3,f4,f5;

f.open("q", ios::out | ios ::app);
f1.open("likelihood", ios::out | ios ::app);
f2.open("p", ios::out | ios ::app);
f3.open("teta", ios::out | ios ::app);
f4.open("likeC", ios::out | ios::app);
f5.open("claster.clu",  ios::out | ios::app);


streamnum = 0;
nstreams = 1;
stream = init_sprng(SPRNG_CMRG, streamnum, nstreams, seed, SPRNG_DEFAULT);


while(!cin.eof())
  {
  cin >>  v1 >> v2;
  cin >> W[v1][v2];
  }



for(i=1;i<=n;i++)
  {
  strenght[i]=0.0;
  for(j=1;j<=n;j++)
    {
    strenght[i]=strenght[i]+W[i][j];
    }
  }


PI=4.0*atan(1.0);
eps=0.0001;
```

```
x=0.0;
for(r=1;r<=G;r++)
  {
  x1=0.0;
  for(j=1;j<=n;j++)
    {
    ksi1=sprng(stream);
    ksi2=sprng(stream);
    teta[r][j]=1.0/n+eps*sqrt(- 2.0*log(ksi1))*cos(2*PI*ksi2);
    x1=x1+teta[r][j];
    }
  for(j=1;j<=n;j++) teta[r][j]=teta[r][j]/x1;
  ksi1=sprng(stream);
  ksi2=sprng(stream);
  pi[r]=1.0/G+eps*sqrt(- 2.0*log(ksi1))*sin(2*PI*ksi2);;
  x=x+pi[r];
  }
for(r=1;r<=G;r++) pi[r]=pi[r]/x;


stop2=0;
while(stop2==0)
{
for(r=1;r<=G;r++)
  {
  for(i=1;i<=n;i++)
    {
    tetao[r][i]=teta[r][i];
    }
  pio[r]=pi[r];
  }
for(r=1;r<=G;r++)
  {
  for(i=1;i<=n;i++)
    {
    for(s=1;s<=G;s++)
      {
      x=0.0;
      epom[s]=0;
      epomc[s]=1;
      stop=0;
      j=1;
      while(stop==0)
```

```
        {
        if(!(tetao[s][j]==0.000000000))
epom[s]=epom[s]+W[i][j]*log(tetao[s][j]);
        else
          {
            if(!(W[i][j]==0))
              {
              stop=1;
              epomc[s]=0;
              }
          }
        if(j==n) stop=1;
        j++;
        }
      if(epomc[s]==1)
                  {
                  if(!(pio[s]==0.0000000000)) epom[s]=epom[s]+log(pio[s]);
                  else epomc[s]=0;
                  }
      }
  if(epomc[r]==1)
     {
     y=0.0;
     for(s=1;s<=G;s++) if(epomc[s]==1) y=y+exp(epom[s]-epom[r]);
     q[r][i]=1.0/y;
     }
  else q[r][i]=0.0;
   }
  }
for(i=1;i<=n;i++)
  {
  x=0.0;
  for(r=1;r<=G;r++) x=x+q[r][i];

  }
x=0.0;
for(r=1;r<=G;r++)
  {
  x1=0.0;
  for(i=1;i<=n;i++)
    {
    x1=x1+q[r][i];
    x2=0.0;
    x3=0.0;
```

```
    for(j=1;j<=n;j++)
      {
      x2=x2+W[j][i]*q[r][j];
      x3=x3+strenght[j]*q[r][j];
      }
    teta[r][i]=x2/x3;
    }
  pi[r]=x1/n;
  x=x+pi[r];
  }

L=0.0;
for (r=1;r<=G;r++)
  {
  for(i=1;i<=n;i++)
    {
    x=0.0;
    x1=0.0;
    for(j=1;j<=n;j++) if(!(teta[r][j]==0) ) x=x+W[i][j]*log(teta[r][j]);
    L=L+q[r][i]*(log(pi[r])+x);
    }
  }

f1 << L << endl;
stop=0;
for(r=1;r<=G;r++)
  {
  for(i=1;i<=n;i++)
    {
    if(fabs(tetao[r][i]-teta[r][i])>control) stop=1;
    }
  if(fabs(pio[r]-pi[r])>control) stop=1;
  }
if(stop==0) stop2=1;
}

for(r=1;r<=G;r++)
  {
  for(j=1;j<=n;j++)
      {
      f << q[r][j] << "\n";
      f3 << teta[r][j] << "\n";
      }
  f  << "\n";
```

```
  f3  << "\n";
  f2 << pi[r] << "\n";
  }

for(r=1;r<=G;r++)
  {
  for(i=1;i<=n;i++)
    {
    if(q[r][i] > 0.5) gindex[i]=r;
    }
  }
f5 << "*Vertices " << n << "\x0d\x0a";
for(i=1;i<=n;i++) f5 << gindex[i]+1 << "\x0d\x0a";

f4 << G << " " << L << "\n";

return EXIT_SUCCESS;

}
```

Literatura

[1] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424:175–308, 2006.

[2] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

[3] József Berke. Spectral fractal dimension. In *TELE-INFO'05: Proceedings of the 4th WSEAS International Conference on Telecommunications and Informatics*, pages 1–4, Stevens Point, Wisconsin, USA, 2005. World Scientific and Engineering Academy and Society (WSEAS).

[4] R. Albert, H. Jeong, and A. L. Barabasi. The diameter of the world wide web. *Nature*, 401:130–131, 1999.

[5] B. Kahng, Y. Park, and H. Jeong. Robustness of the in-degree exponent for the world-wide web. *Phys. Rev. E*, 66(4):046107, Oct 2002.

[6] S. Milgram. The small world problem. *Psychology Today*, 2(1):60–67, 1967.

[7] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks. *Advances in Physics*, 51:1079–1187, jun 2002.

[8] P. Erdős and A. Rényi. On random graphs. I. *Publ. Math. Debrecen*, 6, 1959.

[9] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *ArXiv Condensed Matter e-prints*, oct 1999.

[10] Duncan J. Watts. *Small worlds: the dynamics of networks between order and randomness*. Princeton University Press, Princeton, NJ, USA, 1999.

[11] B. Tadić. Dynamics of directed graphs: the world-wide Web. *Physica A Statistical Mechanics and its Applications*, 293:273–284, 2001.

[12] B. Tadić, G. J. Rodgers, and S. Thurner. Transport on Complex Networks: Flow, Jamming and Optimization. *International Journal of Bifurcation and Chaos*, 17(7):2363–2385, 2007.

[13] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network Motifs: Simple Building Blocks of Complex Networks. *Science*, 298:824–827, 2002.

[14] R. Graben, C. Zhou, M. Thiel, and J. Kurths. *Lectures in Supercomputational Neuroscience: Dynamics in Complex Brain Networks (Understanding Complex Systems.* Springer-Verlafg, Berlin Heidelberg, 2008.

[15] P. R. Villas Boas, F. A. Rodrigues, G. Travieso, and L. da Fontoura Costa. Chain motifs: The tails and handles of complex networks. *Phys. Rev. E*, 77(2):026106–+, 2008.

[16] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proc Natl Acad Sci U S A*, 99(12):7821–7826, 2002.

[17] David Lusseau and M. E. J. Newman. Identifying the role that individual animals play in their social network. *Proc. Soc. B*, 271(477), 2004.

[18] J. Grujić, M. Mitrović, and B. Tadić. Mixing patterns and communities on bipartite graphs of web-based social interactions. *IEEE Xplore*, 2009.

[19] M. Mitrović and B. Tadić. Bloggers behavior and emergent communities in blog space. *Eur. Phys. Journal B*, 73(2):293–301, 2009.

[20] G. Flake, Lawrence, S. Giles, and F. Coetzee. Self-organization and identification of web communities. *Computer*, 35(3), 2002.

[21] T. Jonsson, P. F. Cavanna, D. Zicha, and P. A. Bates. Cluster analysis of networks generated through homology: automatic identification of important protein. *BMC Bioinformatics*, 7, 2006.

[22] R. Guimera and L. A. Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, 2005.

[23] Santo Fortunato and Claudio Castellano. Community structure in graphs. *arXiv:0712.2716*, 2007.

[24] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113, 2004.

[25] J. Reichardt and S. Bornholdt. Statistical mechanics of community detection. *Phys Rev E Stat Nonlin Soft Matter Phys*, 74(1), 2006.

[26] Santo Fortunato, Vito Latora, and Massimo Marchiori. Method to find community structures based on information centrality. *Phys. Rev. E*, 70(5):056104, Nov 2004.

[27] M. Mitrović and B. Tadić. Spectral and dynamical properties in classes of sparse networks with mesoscopic inhomogeneities. *Phys. Rev. E*, 80(2):026123, Aug 2009.

[28] L. Donetti and M. A. Muñoz. Detecting network communities: a new systematic and efficient algorithm. *Journal of Statistical Mechanics: Theory and Experiment*, 10, 2004.

[29] L. Danon, A. Díaz-Guilera, and A. Arenas. The effect of size heterogeneity on community identification in complex networks. *Journal of Statistical Mechanics: Theory and Experiment*, 11, 2006.

[30] A. Arenas, A. Díaz-Guilera, and C. J. Pérez-Vicente. Synchronization Reveals Topological Scales in Complex Networks. *Physical Review Letters*, 96(11):114102–+, 2006.

[31] Haijun Zhou. Distance, dissimilarity index, and network community structure. *Phys. Rev. E*, 67(6):061901, Jun 2003.

[32] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[33] A. Lancichinetti, S. Fortunato, and J. Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015–+, 2009.

[34] M. E. J. Newman and E. A. Leicht. Mixture models and exploratory analysis in networks. *Proceedings of the National Academy of Sciences*, 104(23):9564–9569, 2006.

[35] Roger Guimerà, Marta Sales-Pardo, and Luís A. Nunes Amaral. Modularity from fluctuations in random graphs and complex networks. *Phys. Rev. E*, 70(2):025101, Aug 2004.

[36] J. Huang, T. Zhu, and D. Schuurmans. Web Communities Identification from Random Walks. *Lecture Notes in Computer Science*, 4213:187–198, 2006.

[37] B. Tadić. Exploring Complex Graphs by Random Walks. In *AIP Conf. Proc. 661: Modeling of Complex Systems*, 2003.

[38] K.-I. Goh, B. Kahng, and D. Kim. Spectra and eigenvectors of scale-free networks. *Phys. Rev. E*, 64(5):051903, 2001.

[39] A. Banerjee and J. Jost. Spectral plot properties: towards a qualitattive classification of networks. *Networks and Heterogenieous Media*, 3(2):395–411, 2008.

[40] M. Mitrović and B. Tadić. Search of Weighted Subgraphs on Complex Networks with Maximum Likelihood Methods. *Lecture Notes in Computer Science*, 5102:551–558, 2008.

[41] IMDb, 1990.

[42] B92. Blog B92. *http://blog.b92.net*, 2007.

[43] V. Batagelj and A. Mrvar. pajek, 1996.

[44] M. E. J. Newman. Analysis of weighted networks. *Phys. Rev. E*, 70(5):056131, Nov 2004.

[45] J. Živković, B. Tadić, N. Wick, and S. Thurner. Statistical indicators of collective behavior and functional clusters in gene networks of yeast. *European Physical Journal B*, 50:255–258, 2006.

[46] J. Živković, M. Mitrović, and B. Tadić. Correlation patterns in gene expressions along the cell cycle of yeast. *Studies in comutational intelligence*, 207, 2009.

[47] J. Grujić. Movies recommendation networks as bipartite graphs. *Lecture Notes in Computer Science*, 5102, 2008.

[48] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, and Gabrielian. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell*, 2(1), 1998.

[49] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. *Nature genetics*, 22(3):281–285, 1999.

[50] A. Madi, Y. Friedman, D. Roth, T. Regev, S. Bransburg-Zabary, and E.B. Jacob. Genome holography: Deciphering function-form motifs from gene expression data. *PLoS ONE*, 3(7):e2708, Jul 2008.

[51] B. Tadić and M. Mitrović. Jamming and correlation patterns in traffic of information on sparse modular networks. *European Physical Journal B*, 71:631–640, October 2009.

[52] S. Wasserman, K. Faust, and D. Iacobucci. *Social Network Analysis : Methods and Applications (Structural Analysis in the Social Sciences)*. Cambridge University Press, November 1994.