Analiza socijalnih mreža

Algoritmi za detekciju komuna u mreži

Marko Mišić, Jelica Protić

13M111ASM

2017/2018.

Tehnike za detekciju komuna (1)

- *Top-down* pristupi
  - Polaze od kompletne mreže i pokušavaju da izvrše hijerarhijsku podelu odozgo na dole
  - Najčešće kao rezultat daju dendogram koji prikazuje grupisanje manjih jedinica u veće
  - Analiza povezanih komponenti, *Girvan-Newman* metod
- *Bottom-up* pristupi
  - Grade veće grupe polazeći od manjih gradivnih jedinica odozdo na gore
  - Propagacija labela, *Louvain* metod i sl.

Tehnike za detekciju komuna (2)

- Particijsko grupisanje
  - *k-means clustering* i slične metode
- Hijerarhijska klasterizacija
  - Aglomerativni algoritmi
    - *Ravasz* algoritam
  - Razorni algoritmi
    - *Girwan-Newman* metoda
- Propagacija labela
  - Sinhrona, semisingrona i asinhrona propagacija
  - Balansirana propagacija
- Koristi se pretpostavka da je graf neusmeren

Particijsko grupisanje (1)

- Najstarija tehnika
  za podelu mreže na komune (klastere)
  - Pristup potiče iz statističke analize podataka
- Tehnika se oslanja na podelu čvorova
  izabranog grafa u *k* grupa predefinisanih veličina
  - Grupe su takve
    da je broj grana između grupa minimalan
- Neophodno je specificirati i
  broj grupa i njihovu veličinu
- Veći broj algoritama

Particijsko grupisanje (2)

- Problem trivijalnih podela čvorova na grupe
  - Ukoliko se ne odredi unapred broj grupa, rezultujuća podela grupiše sve čvorove u jednu grupu
    - Traži samo minimalan broj grana između njih
    - Broj grana između grupa će tada biti jednak nuli
  - Ukoliko se ne specificira željena veličina grupa, rezultujuća podela daje dve grupe
    - Minimalan broj grana između dobijenih grupa se dobija odvajanjem čvora sa najnižim stepenom u posebnu komunu u odnosu na ostatak grafa

Particijsko grupisanje (3)

- Najpre se odredi broj grupa *k*
- Čvorovi su prikazani kao tačke u prostoru
  - Rastojanje između svaka dva čvora predstavlja njihovu udaljenost
  - Udaljenost čvorova se koristi kao mera njihove sličnosti (razlike)
- Bira se funkcija cene nad kojom se vrši minimizacija ili maksimizacija podelom čvorova na *k* grupa
  - Ona koristi izračunate razdaljine između čvorova ili udaljenost čvorova grafa od određenih tačaka u prostoru (centroida)

Particijsko grupisanje (4)

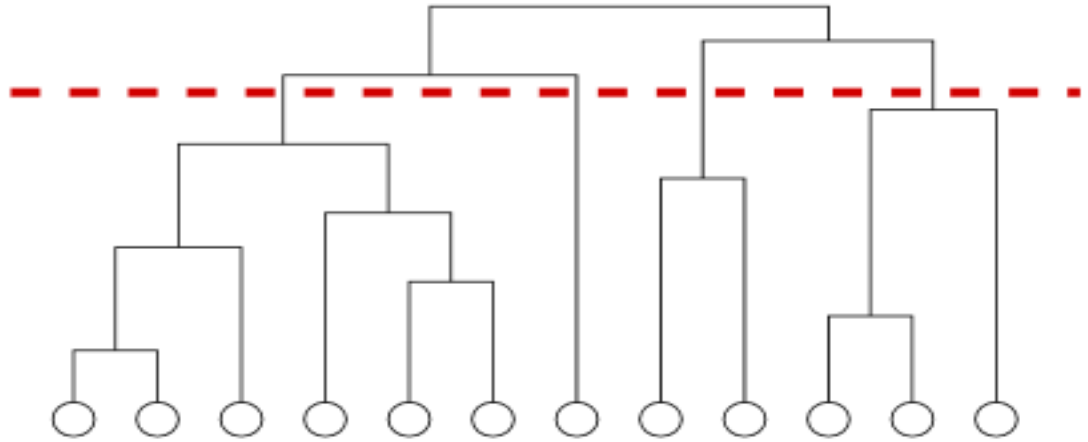
- Metodi koji ne koriste centroide
  - *minimum k-clustering* (minimizacija dijametra)
  - *k-clustering sum*
    (minimizacija srednje vrednosti udaljenosti)
- Metodi koji koriste centroide
  - *k-center, k-median* (koriste udaljenost od centroida)
  - *k-means clustering*
    (minimizacija sume kvadrata udaljenosti)
    - Najčešće korišćena metoda iz grupe

Hijerarhijski metodi (1)

- Često nije moguće unapred poznavati karakteristike mreže i njenih komuna
  - Ne mogu se uvoditi pretpostavke na osnovu kojih bi se izvršila klasterizacija
    - Karakteristično za particijske metode
- Hijerarhijski metodi dele mrežu na komune bez ikakvih pretpostavki o strukturi mreže
  - Oslanjaju na premisu da je moguće inkrementalnim grupisanjem jako povezanih čvorova u grafu podeliti graf na komune
  - Potrebno je odrediti meru za jačinu povezanosti ili sličnost čvorova na osnovu koje se vrši grupisanje
  - Na osnovu korišćenih mera sličnosti se razlikuju metodi
  - Formira se matrica sličnosti za sve parove čvorova mreže

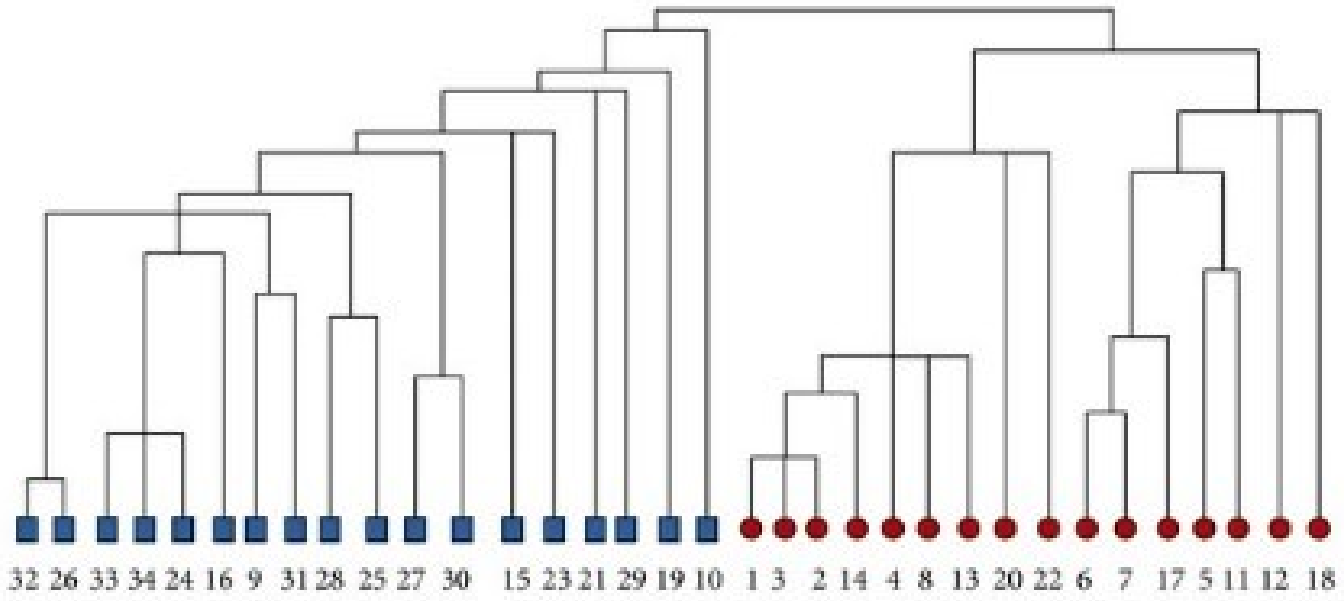
Hijerarhijski metodi (2)

- Rezultati hijerarhijskog grupisanja mogu se prikazati pomoću dendrograma
  - Hijerarhijsko stablo
  - Idealno za ilustraciju rasporeda klastera
  - Analizom horizontalnih preseka stabla može se uočiti napredak podele grafa na komune u različitim fazama obrade



Hijerarhijski metodi (3)

- *Zachary's Karate Club* – primer dendograma
  - Čvor 1 je predsednik, a 34 instruktor



Aglomerativni i razorni algoritmi

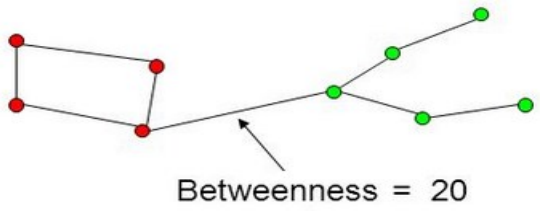
- Aglomerativni algoritmi pretpostavljaju da je svaki čvor komuna za sebe
  - Grupiše čvorove na osnovu njihove međusobne sličnosti
  - Na kraju postupka je ceo graf jedna komuna
- Razorni algoritmi posmatraju ceo graf kao jednu komunu
  - Dele mrežu na manje komune razdvajanjem grupa čvorova na osnovu njihovog razlikovanja ili nedostatka sličnosti
  - Postupak se završava podelom grafa na broj komuna koji je jednak broju čvorova
  - Svaki čvor predstavlja komunu za sebe
- Uvode se kriterijumi zaustavljanja da bi se izabrala neka podela
  - Eksplicitan broj klastera ili funkcija koja ocenjuje kvalitet postignute podele

Girvan-Newman metod (1)

- Najpopularniji razorni algoritam
  - Polazi od početnog grafa kao jedne komune
  - Formira komuna iterativnim uklanjanjem određenih grana iz grafa
- Četiri osnovna koraka u radu algoritma
  - Računanje centralnosti svake grane u skupu grana mreže
    - *Edge betweenness*
  - Uklanjanje grane koja ima najveću centralnost
  - Ponovo računanje centralnosti grana u novom, modifikovanom grafu dobijenom u prethodnom koraku
  - Iterativno ponavljanje prethodnih koraka dokle god postoje grane u tekućem grafu

Girvan-Newman metod (2)

- Različite tehnike za računanje centralnosti grane
  - Geodezijska, *random-walk* i *current-flow* centralnost
- Najčešće korišćena geodezijska (relaciona) centralnost grane
  - Grane koje spajaju komune se nalaze na velikom broju puteva koji spajaju čvorove različitih komuna
    - Biće pre uklonjene iz grafa
  - Predstavlja meru uticaja grane na prostiranje informacija kroz mrežu
  - Računa se kao broj najkraćih puteva između svaka dva čvora u grafu na kojima se data grana nalazi
    - Ukoliko između dva čvora ima više puteva najkraće dužine, njima se dodeljuju težine tako da im je zbir težina jednak jedinici



- Velika složenost algoritma ~ $O(mn^2)$
  - Neprimenljiv za grafove sa više od ~10-20 hiljada čvorova

Mera kvaliteta podele (1)

- Mere kvaliteta kvantifikuju valjanost podele
  - Pomažu u odabiru adekvatne podele
  - Omogućavaju da se svakoj podeli grafa na komune dodeli broj kao oznaku kvaliteta te podele
  - Podele se rangiraju i bira se najbolja
- Ista mera kvaliteta se mora primeniti nad svim podelama
  - Računanje na nivou čvora, komune ili celog grafa
- *Girvan-Newman* modularnost
  - Računa se u odnosu na prazan (*null*) model grafa
    - Graf sa slučajnim rasporedom grana koji ne sadrži komune
    - Poređenjem gustine grana u komuni sa gustinom grana u podgrafu istih čvorova u praznom modelu dobija se odstupanje strukture komune od slučajnog rasporeda grana

Mera kvaliteta podele (2)

- Računa se po sledećoj formuli:

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(C_i, C_j)$$

  - Suma se računa za svaki par čvorova *i* i *j*
  - A_{ij} predstavlja broj grana između čvorova *i* i *j*
  - k_i i k_j predstavljaju stepene čvorova *i* i *j* u praznom modelu, a čitav proizvod verovatnoću postojanja grane
  - Ukoliko *i* i *j* pripadaju istoj komuni $\delta(C_i, C_j)$ je jednaka jedan, a u suprotnom nula
- U sumi učestvuju samo oni parovi čvorova koji pripadaju istim komunama:

$$Q = \sum_{c=1}^{n_c}\left[\frac{l_c}{m} - \left(\frac{d_c}{2m}\right)^2\right]$$

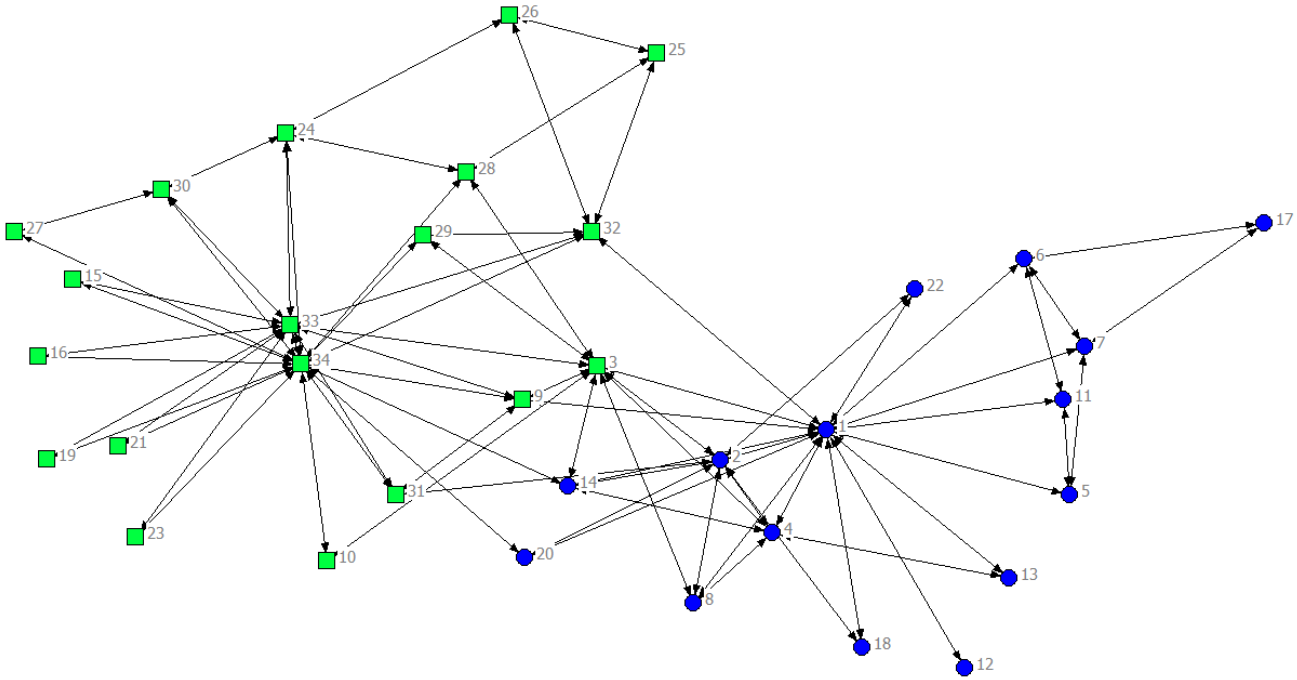
  - gde je n_c broj komuna, l_c je broj grana unutar komune
  - d_c zbir stepeni čvorova unutar komune

Mera kvaliteta podele (3)

- *Girvan-Newman* modularnost uzima vrednosti u opsegu [-1/2, 1)
  - Za komunu u kojoj se nalazi ceo graf, vrednost modularnosti je nula
  - U praksi se najčešće dobijaju vrednosti funkcije u rasponu od 0.3 do 0.7
  - Podela grafa koja ima modularnost od 0.7 se smatra veoma dobrom podelom
- Alternativna metrika - konduktansa (*conductance*)
  - Kvalitet procenjuje na osnovu spoljašnje povezanosti komune
  - Predstavlja odnos broja grana koje izlaze iz komune u ostatak grafa i broja grana koje imaju bar jedan čvor unutar komune

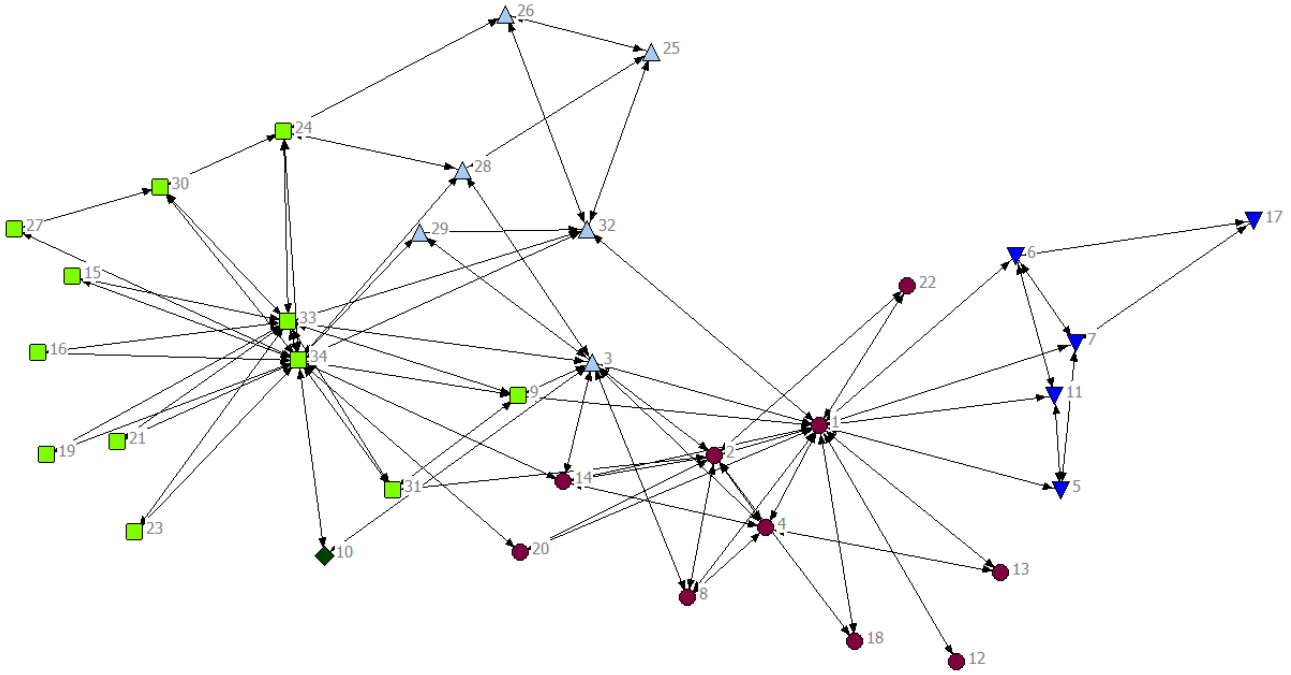
Mera kvaliteta podele (4)

- *Zachary's karate club* mreža
  - Podeljena na dve komune *Girvan-Newman* metodom
  - Modularnost ove podele iznosi Q = 0.360



Mera kvaliteta podele (5)

- *Zachary's karate club* mreža
  - Podeljena na pet komuna *Girvan-Newman* metodom
  - Modularnost ove podele iznosi Q = 0.401



Propagacija labela (1)

- Propagacijom labela je metod koji podelu grafa na komune vrši u skoro linearnom vremenu
  - Komune se dobijaju kao grupe čvorova koji dele istu vrednost labele
- Nastao iz potrebe da se u masivnim grafovima, na brz i efikasan način detektuju komune
  - Efikasno radi i sa više miliona čvorova
  - Obrada realni mreža
- Ne zahteva nikakve dodatne informacije i pretpostavke o grafu
  - Samo rasporeda čvorova i grana
- Ne dobija se uvek jedinstveno rešenje
  - Koristi dodatne tehnike za odabir najbolje particije

Propagacija labela (2)

- Propagacija labela se vrši na sledeći način:
  - Na početku se svakom čvoru dodeli jedinstvena labela
  - Potom se nasumičnim redosledom obilaze svi čvorovi grafa i nad njima vrši preračunavanje vrednosti labela
  - Svakom čvoru se dodeljuje vrednost labele koju dele najveći broj njegovih suseda
    - Ukoliko ne postoji jedna jedinstvena vrednost koja je maksimalno zastupljena u susedstvu čvora, nasumično se bira jedna od vrednosti koje dele maksimalnu zastupljenost
  - Proces obilaska i preračunavanja se ponavlja do konvergencije
  - Konvergencija je postignuta u trenutku kada svaki čvor ima vrednost labele koja je maksimalno zastupljena u njegovom susedstvu

Propagacija labela (3)

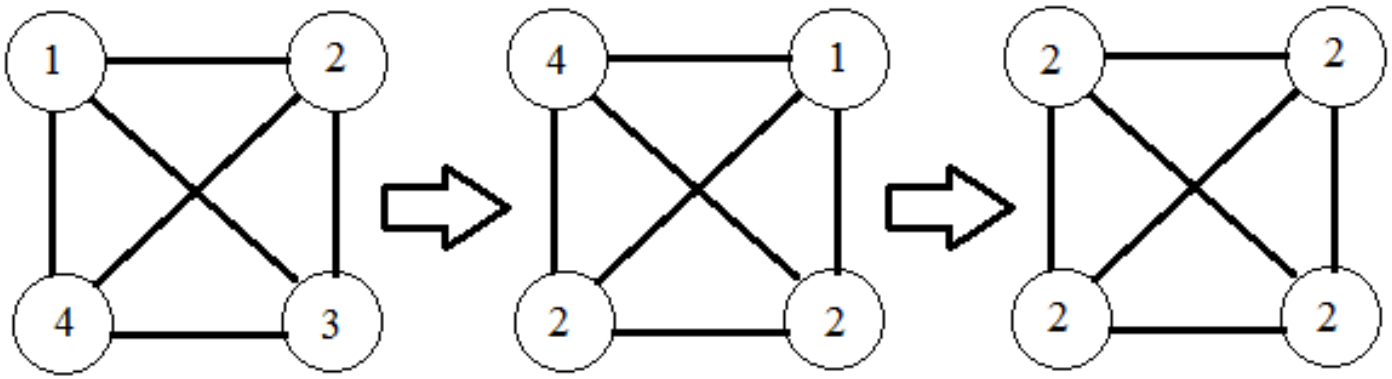
- Usvajanje labela se vrši po sledećoj formuli:

$$c_v = argmax_l \sum_{w \in \mathcal{N}^l(v)} w(v,w)$$

  - c_v je komuna kojoj pripada čvor v
  - Vrednost $w(v, w)$ je jedan u slučaju netežiskog grafa ili je jednako težini grane između čvorova v i w
  - Čvor w pripada skupu $\mathcal{N}^l(v)$ koji čine susedi čvora v sa labelom l
- Čvorovi unutar komune su dosta bolje povezani međusobno nego sa ostatkom grafa
  - Većina čvorova iz iste komune će deliti labelu već posle nekoliko iteracija
- Kada čvorovi postignu stabilno stanje, čvorovi sa istom labelom predstavljaju jednu komunu

Propagacija labela (4)

- Primer rada jedne varijacije osnovnog algoritma
  - Svaki čvor razmatra labele suseda iz prethodne iteracije
    - Sinhorna propagacija
  - U nerešenim situacijama se uzima nasumično odabrana labela iz skupa najfrekventnijih ili zadržava tekuća ukoliko je i ona u skupu najfrekventnijih labela
    - Prva iteracija
    - Donji čvorovi u poslednjoj iteraciji

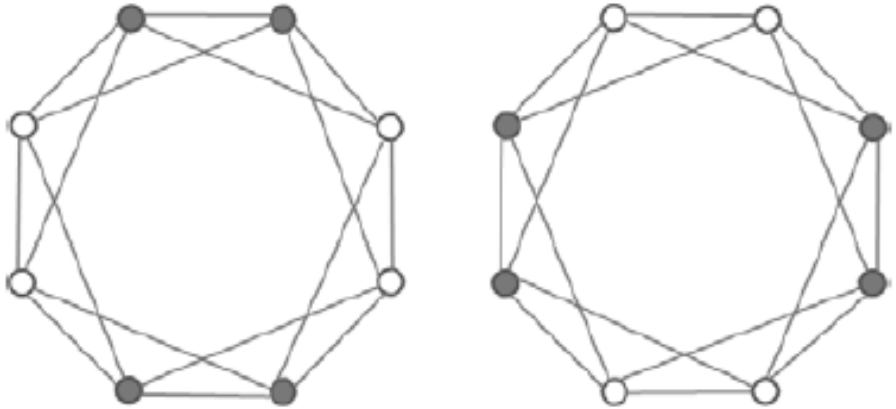


Problemi konvergencije

- Za neke tipove mreže,
  originalni algoritam se nikad ne zaustavlja
  - Dešavaju se oscilacije labela po iteracijama
- Na zaustavljanje algoritma imaju uticaja:
  - Način ažuriranja labela (tip propagacije)
    - Koje informacije o labelama se koriste u trenutku ažuriranja?
  - Strategija razrešavanja labela
    sa istom maksimalnom frekvencijom pojavljivanja u komuni
    - Da li se bira na slučajan način ili drugačije?
  - Kriterijum zaustavljanja
    - Koliko dugo se mora održati ekvilibrijum pre zaustavljanja?

Tip propagacije (1)

- Sinhrona propagacija
  - Svaki čvor u tekućoj iteraciji računa novu labelu na osnovu labela suseda iz prethodne iteracije
  - Problem oscilacija labela
- Primer sa slike
  - Jedan sused svoje boje
  - Tri suseda suprotne boje
  - Nikad se ne dostiže stabilno stanje



- U opštem slučaju i duži period oscilovanja

Tip propagacije (2)

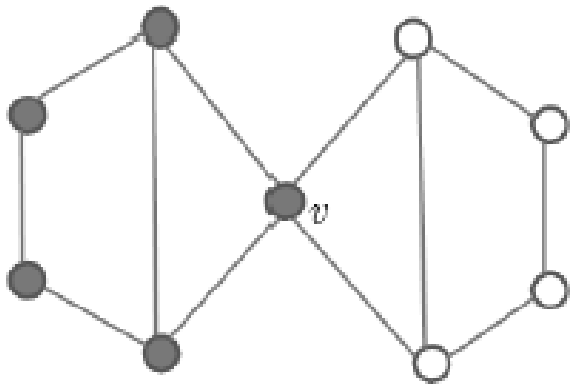
- Asinhrona propagacija
  - Čvorovi se ažuriraju sekvencijalno u nekom slučajnom poretku
  - Poredak je drugačiji od iteracije do iteracije

$$C_x(t) = f\left(C_{x_{i1}}(t), \ldots, C_{x_{im}}(t), C_{x_{i(m+1)}}(t-1), \ldots, C_{x_{ik}}(t-1)\right)$$

  - Susedi čvora x koji su ažurirani u tekućoj iteraciji su x_{i1}, ..., x_{im}, dok čvorovi $x_{i(m+1)}$, ...,x_{ik} još uvek nisu ažurirani u tekućoj iteraciji
  - Problem stabilnosti rešenja
    - Posledica slučajnog poretka za ažuriranje
- Semisinhrona propagacija
  - Zasniva se na bojenju grafa tako da susedi imaju različite boje
    - Vremenska složenost bojenja
  - U svakoj fazi čvorovi iste boje ažuriraju svoje labele
  - Stabilan i brz

Strategija razrešavanja labela (1)

- Problem odabira nove labele čvora u situaciji kada dve ili više različitih labela ispunjavaju kriterijum za propagaciju
  - Nerešena situacija (*tie*)
  - Utiče na kvalitet rešenja
  - Sprečava pojavu oscilacija labela



Strategija razrešavanja labela (2)

- Nekoliko tipičnih pristupa
- Slučajan odabir (*LPA-Rand*)
  - Osnovna varijanta
- Davanje prednosti tekućoj labeli (*LPA-Prec*)
  - *Label retention*
  - Dovodi do brže konvergencije algoritma,
    ali particije mogu biti slabijeg kvaliteta nego kod *LPA-Rand*
- Uzima se labela sa najvećim prioritetom (*LPA-Max*)
  - Ako su labele celi brojevi,
    najveći prioritet može imati najveći ili najmanji broj
- Kombinacija prethodna dva pristupa (*LPA-Prec-Max*)
  - Gleda se prioritet, ali se prednost daje tekućoj labeli

Kriterijum zaustavljanja

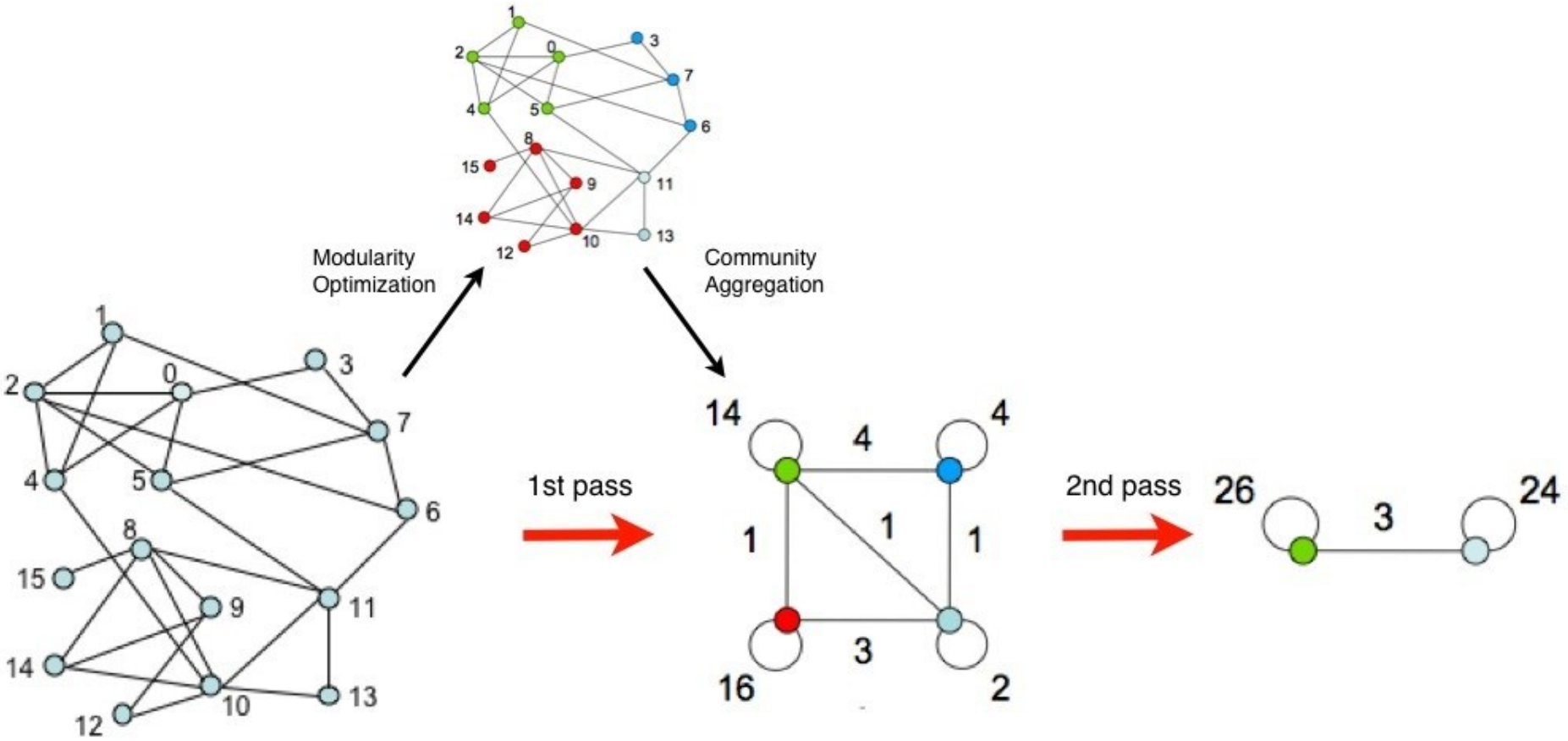
- Nakon nekoliko iteracija algoritma,
  većina čvorova usvaja konačnu vrednost svoje labele
  - Problem su određene konfiguracije grafa
- Različiti kriterijumi zaustavljanja
  kako bi se sprečilo beskonačno izvršavanje:
  - Kada nastupi iteracija u kojoj svaki čvor deli istu labelu
    kao i većina njegovih suseda
  - Može da ne nastupi nikada, zbog oscilacija
  - Svaki čvor ima istu labelu
    kao u prošloj ili pretprošloj iteraciji
    uz korišćenje *LPA-Max* strategije za razrešavanje labela
    - Kod LPA-Max su moguće oscilacije samo sa periodom 2
  - Istekne određeni broj iteracija

Louvain metod (1)

- Metod zasnovan na maksimizaciji modularnosti particija u okviru mreže
  - *NP*-hard problem za tačno rešenje
  - Pohlepan algoritam, prosečne složenosti ~ *O(nlogn)*
- Algoritam se izvršava u dva koraka
  - Najpre se pronalaze manje komune i optimizuje modularnost lokalno
  - Proces se ponavlja sve dok postoji uvećanje modularnosti
  - Zatim se agregiraju se čvorovi koji pripadaju istim komunama i pravi mreža „komuna"
  - Postupak se ponavlja od prvog korakanad dobijenom mrežom dok se ne dostigne maksimalna modularnost
- Izlaz je više podela različite granularnosti

Louvain metod (2)

- Primer rada algoritma



Literatura

- Santo Fortunato, Community detection in graphs", Physics Reports, 2010.
- L. Šubelj, Label propagation for partitioning, Advances in Network Clustering and Blockmodeling, 2017.
- http://www.network-science.org/
- https://www.pulsarplatform.com/blog/2014/detecting-communities-using-social-network-analysis/