

SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO–MATEMATIČKI FAKULTET
MATEMATIČKI ODSJEK

Adrian Satja Kurdija

**GRAFOVI PREFERENCIJE I NJIHOVA
PRIMJENA**

Diplomski rad

Voditelj rada:
doc. dr. sc. Lavoslav Čaklović

Zagreb, 2015.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Ponajviše zahvaljujem svojim roditeljima na neizmjenoj podršci tijekom studija.

Iskreno zahvaljujem svojem mentoru, doc. dr. sc. Lavoslavu Čakloviću koji me je zainteresirao za ovo područje, ukazao povjerenje te pružio veliku podršku i pomoć.

Sadržaj

Sadržaj	iv
Uvod	2
1 Graf preferencije i njegov potencijal	3
1.1 Graf preferencije	3
1.2 Potencijal	5
1.3 Svojstva potencijala	9
2 Agregacija grafova preferencije	12
2.1 Grupni (multi)graf preferencije	12
2.2 Grupni potencijal	14
3 Univerzalna izborna procedura	18
3.1 Pretvorba glasačkih listića u grafove preferencije	19
3.2 Eksperiment	25
4 Klasterizacija grafova preferencije	28
4.1 Klasterizacija skupine na temelju preferencija njezinih članova	29
4.2 Hijerarhijska klasterizacija	30
4.3 Particijska klasterizacija metodom k -sredina	32
4.4 Eksperiment	33
5 Dodatak	37
Bibliografija	41

Uvod

U situacijama gdje je potrebno rangirati ili evaluirati neke objekte, oni se često uspoređuju u parovima, osobito ako precizno mjerenje nije moguće. U mnogim eksperimentima, osobito u društvenim znanostima, rangiranje objekata (npr. umjetničkih djela) vrši se na način da se ispitaniku objekti pokazuju u parovima, a ispitanik odgovara koji mu se od prikazanih dvaju objekata više sviđa [6]. Ljudima je mnogo lakše uspoređivati, nego evaluirati [8], [19]. Na osnovi tih usporedbi izvode se daljnji zaključci u raznim modelima kao što su *law of comparative judgment* u psihologiji [17], ili *analytic hierarchy process* (AHP) u teoriji odlučivanja [14]. Prvi koji je eksplicitno primijenio usporedbu parova u izornoj proceduri bio je Condorcet [5].

U ovom radu fokusirat ćemo se na *graf preferencije* kao osnovnu strukturu dobivenu iz preferencija izraženih usporedbama. Grubo rečeno, vrhovi grafa preferencije predstavljaju alternative koje se uspoređuju, a usmjereni bridovi opisuju usporedbe. Metoda potencijala koju je razvio L. Čaklović [22] omogućit će nam da iz toga grafa dobijemo vrijednosti alternativa, a stoga i njihovo rangiranje. Nadalje, poopćenje metode potencijala na *preferencijski multigraf*, što je originalna autorova zamisao, omogućit će agregaciju preferencija više osoba (koji često nisu međusobno suglasni), kao i klastersku analizu (grupiranje) skupine osoba na temelju njihovih preferencija.

U točkama koje slijede navodimo neke situacije u kojima se javljaju grafovi preferencije, tj. u kojima su primjenjive neke od spomenutih metoda.

– *Društvene znanosti*. U znanostima poput psihologije ili sociologije, grafovi preferencije implicitno se javljaju u raznim anketnim istraživanjima gdje se odgovarajući upitnik temelji na izražavanju preferencija. U tom kontekstu može biti potrebna i klasterska analiza skupine ispitanika ili agregacija njihovih preferencija.

– *Specifične društvene skupine*. Članovi nekih društvenih skupina izražavaju preferencije kao dio svojega posla – npr. suci, komisije, povjerenstva, odbori i osobito parlament. Takvu skupinu moguće je klasterski analizirati ili agregirati njihove preferencije.

– *Višekriterijsko odlučivanje*. U višekriterijskom (višeatributnom) odlučivanju postoji

određen broj kriterija, a svaki kriterij daje nam neke informacije o tome koje su alternative bolje ili lošije [21]. Primjerice, ako je cilj kupnja automobila te razmatramo n alternativa (automobila), kriteriji mogu biti cijena, veličina, udobnost, izgled, potrošnja goriva i tako dalje. Ako kriterij shvatimo kao objekt koji izražava preferencije (npr. kriterij "udobnost" kaže da je automobil A bolji od automobila B), možemo izvršiti klustersku analizu kriterija (s ciljem da kriterije koji daju slične preferencije grupiramo zajedno) ili agregirati njihove preferencije da bismo donijeli konačnu odluku.

– *Sportovi*. U mnogim sportovima sudionici se sparuju u mečeve (utakmice) gdje se natječu jedan protiv drugoga. Svaki od tih mečeva možemo gledati kao usporedbu para, čime dobivamo graf preferencije iz kojega možemo izvući rangiranje.

– *Pretraživanje dokumenata (engl. document retrieval)*. Pretraživač ili dio pretraživača za određeni upit pronalazi više odnosno manje relevantne (preferirane) dokumente i tako daje graf preferencije koji možemo koristiti za rangiranje dokumenata, tj. rezultata pretrage [12]. U slučaju više dijelova pretraživača ili više samih pretraživača, njihove preferencije valja agregirati. (Pretraživač je nešto općenitiji pojam od web-tražilice.)

– *Suradničko filtriranje (engl. collaborative filtering)*. U sustavima za predviđanje ili preporuke (engl. *recommender systems*) važno je predvidjeti kvalitetu određene usluge za određenog korisnika. U tom kontekstu [7] predlaže agregaciju korisničkih grafova preferencije. Za predikcije kvalitete usluge koriste se sličnosti između korisnika (ili usluga) pa se i njihova klusterska analiza prirodno nameće [24].

– *Strojno učenje*. U kontekstu tzv. *preference learninga*, [12] koristi graf preferencije kao dio kernel metode za učenje funkcije vrednovanja

– *Izborne procedure*. O njima ćemo više reći u Poglavlju 3.

Još neke primjene agregacije, odnosno klasterizacije grafova preferencije bit će spomenute u Poglavljima 2 i 4 koja se bave tim temama.

Rad je organiziran na sljedeći način. Poglavlje 1 obrađuje ključne pojmove: graf preferencije i metodu potencijala. Poglavlje 2 govori o agregaciji grafova preferencije, ostvarenoj poopćenjem metode potencijala na preferencijski multigraf. Primjena te agregacije na izbornu proceduru obrađena je u Poglavlju 3. Spomenuta Poglavlja 2 i 3 temeljena su na vrlo nedavnom istraživačkom radu L. Čaklovića i autora [23]. Druga važna primjena (poopćene) metode potencijala je već spomenuta klusterska analiza koja je detaljno obrađena u Poglavlju 4, implementirana u programskom jeziku Python (vidi Dodatak) i eksperimentalno primijenjena na zemlje sudionice Eurosong natjecanja na temelju glasačkih preferencija. To je poglavlje autorov doprinos.

Poglavlje 1

Graf preferencije i njegov potencijal

Ovo poglavlje opisuje graf preferencije i metodu potencijala (engl. Potential Method), razvijenu od L. Čaklovića i detaljnije obradenu npr. u [21]. Graf preferencije opisuje preferencije jednog aktera (npr. glasača, ispitanika, kriterija...) nad određenim skupom alternativa, a iz *potencijala* grafa preferencije iščitavamo implicitne vrijednosti alternativa inducirane tim preferencijama.

1.1 Graf preferencije

Neka je V skup alternativa nad kojima izražavamo preferencije. Ako alternativu u preferiramo nad alternativom v , u oznaci $u > v$, to možemo prikazati usmjerenim bridom koji izlazi iz vrha v te ulazi u vrh u na grafu čiji smo skup vrhova poistovjetili sa skupom alternativa V . Taj brid označavamo kao uređeni par (u, v) , tj. prva je komponenta ulazni, a druga izlazni vrh. Ako je preferencija dodatno opisana svojim intenzitetom na određenoj skali (npr. jednaka, slaba, umjerena, izrazita ili apsolutna preferencija), što je moguće izraziti nenegativnim realnim brojem (npr. 0, 1, 2, 3 ili 4), tada usmjerenom bridu iz v u u pridajemo i težinu koju zovemo *tokom preferencije* i označavamo kao $\mathcal{F}_{(u,v)}$. U slučaju jednake preferencije alternativa u i v (u oznaci $u \sim v$), vrijedi $\mathcal{F}_{(u,v)} = 0$, a smjer brida nije važan, tj. brid (u, v) možemo zamijeniti bridom (v, u) . Ovime smo motivirali sljedeću definiciju.

Definicija 1.1. *Graf preferencije uređena je trojka $\mathcal{G} = (V, \mathcal{A}, \mathcal{F})$ gdje je V skup od $n \in \mathbb{N}$ vrhova, $\mathcal{A} \subseteq V \times V$ skup od $m \in \mathbb{N}_0$ usmjerenih bridova, a $\mathcal{F} : \mathcal{A} \rightarrow \mathbb{R}_{\geq 0}$ funkcija koja svakom bridu (u, v) pridružuje odgovarajući intenzitet preferencije $\mathcal{F}_{(u,v)}$ i nazivamo je tokom.*

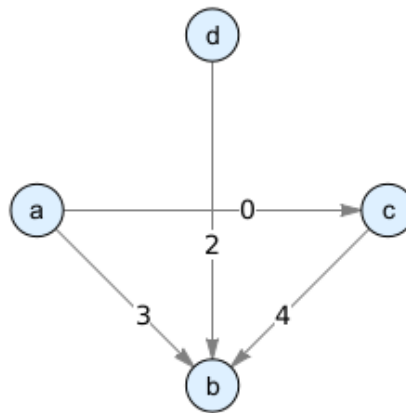
Primjer 1.2. *Pretpostavimo da je osoba dala sljedeće iskaze o alternativama a, b, c i d :*

- b je za 3 bolja od a ,
- c je jednako dobra kao a ,
- b je za 4 bolja od c ,
- b je za 2 bolja od d .

Dakle, preferencije ove osobe su $b > a$, $c \sim a$, $b > c$ i $b > d$ sa intenzitetima 3, 0, 4 i 2, redom. U odgovarajućem grafu preferencije postojat će dakle sljedeći bridovi:

- brid (b, a) s tokom $\mathcal{F}_{(b,a)} = 3$,
- brid (c, a) s tokom $\mathcal{F}_{(c,a)} = 0$,
- brid (b, c) s tokom $\mathcal{F}_{(b,c)} = 4$,
- brid (b, d) s tokom $\mathcal{F}_{(b,d)} = 2$.

Graf je prikazan na Slici 1.1.



Slika 1.1: Graf koji opisuje preferencije iskazane u Primjeru 1.2.

Stupanj vrha v označavamo kao $\deg(v)$ i on je jednak broju bridova koji ulaze ili izlaze iz v . U gornjem je primjeru $\deg(a) = \deg(c) = 2$, $\deg(d) = 1$ i $\deg(b) = 3$.

$m \times n$ matricu $A = (A_{\alpha,v})$ zovemo *matricom incidencije* grafa i definiramo je na sljedeći način:

$$A_{\alpha,v} = \begin{cases} -1, & \text{ako brid } \alpha \text{ izlazi iz vrha } v, \\ 1, & \text{ako brid } \alpha \text{ ulazi u vrh } v, \\ 0, & \text{inače,} \end{cases} \quad (1.1)$$

pri čemu pretpostavljamo da su skup vrhova i skup bridova numerirani, tj. poistovjećujemo ih sa skupovima $\{1, \dots, n\}$ i $\{1, \dots, m\}$. Iz ove pretpostavke slijedi i da tok \mathcal{F} možemo poistovjetiti s vektorom iz \mathbb{R}^m , što ćemo u nastavku i činiti.

$n \times n$ matricu $L = A^T A$ zovemo *Laplaceovom matricom* grafa i lako se pokazuje da vrijedi:

$$L_{ij} = \begin{cases} -1, & \text{ako postoji brid } (i, j) \text{ ili } (j, i), \\ \deg(i), & \text{ako je } i = j, \\ 0, & \text{inače.} \end{cases} \quad (1.2)$$

Za graf iz gornjeg primjera (Slika 1.1) navedene matrice izgledaju ovako:

$$A = \begin{bmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}, \quad L = \begin{bmatrix} 2 & -1 & -1 & 0 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 2 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix}.$$

1.2 Potencijal

Cilj je metode potencijala na temelju grafa preferencije odrediti vrijednosti alternativa (vrhova) tako da one odgovaraju intenzitetima preferencija, tj. toku grafa. Na primjer, ako je $u > v$ s intenzitetom 5, želimo da vrijednost alternative u bude za 5 veća od vrijednosti alternative v . Te vrijednosti čine vektor $X \in \mathbb{R}^n$ koji zovemo *potencijalom* grafa preferencije, pri čemu je komponenta X_v vrijednost alternative v . Kao što je naznačeno, želimo da potencijal zadovoljava

$$X_u - X_v = \mathcal{F}_{(u,v)}, \quad \forall (u, v) \in \mathcal{A}. \quad (1.3)$$

Ovo se, naravno, svodi na sustav jednadžbi. Da bismo sustav lijepo zapisali, primijetimo da je $X_u - X_v$ jednak skalarnom produktu vektora X i onog retka matrice incidencije A koji odgovara bridu (u, v) . Iz toga slijedi da sustav jednadžbi glasi

$$AX = \mathcal{F}. \quad (1.4)$$

Nažalost, ovaj sustav ne mora imati rješenje. Na primjer, ako graf sadrži ciklus generiran usporedbama $a > b, b > c, c > a$ s intenzitetima 1, moralo bi vrijediti $X_a > X_b, X_b > X_c$

i $X_c > X_a$, što nije moguće. Međutim, takve preferencije očito nisu konzistentne, pa se postavlja pitanje treba li ih uopće dopustiti.

Primijetimo da postoje i suptilniji primjeri nekonzistentnih preferencija. Na primjer, $a > b$, $b > c$ i $a > c$ očito su konzistentne preferencije, ali i intenziteti bi se morali podudarati: tok $\mathcal{F}_{(a,c)}$ morao bi biti jednak zbroju tokova $\mathcal{F}_{(a,b)}$ i $\mathcal{F}_{(b,c)}$. Drugi je primjer graf sa Slike 1.1: budući da je $\mathcal{F}_{a,c} = 0$ (tj. $a \sim c$), očekujemo da se a i c jednako razlikuju od alternative b , što nije u skladu sa $\mathcal{F}_{(b,a)} = 3$ i $\mathcal{F}_{(b,c)} = 4$. Dovoljno je izjednačiti ove dvije vrijednosti toka da graf postane konzistentan.

U takvim i sličnim, a pogotovo malo složenijim slučajevima, zahtijevati potpunu konzistentnost ispitanika nije realno. Zato ne postavljamo *nikakav* zahtjev za konzistentnošću iskazanih preferencija pa sustav (1.4) ne rješavamo točno (ako nema rješenja), nego približno, kako će biti opisano u nastavku teksta.

Gornja diskusija motivira sljedeću definiciju:

Definicija 1.3. *Graf preferencije je konzistentan ako postoji rješenje odgovarajućeg sustava $AX = \mathcal{F}$.*

Potencijal nekonzistentnog grafa dobit ćemo kao najbolje približno rješenje sustava (1.4) u smislu najmanjih kvadrata. Tražimo dakle

$$\arg \min_{X \in \mathbb{R}^n} \|AX - \mathcal{F}\|. \quad (1.5)$$

Odgovarajući AX je, dakle, vektoru \mathcal{F} najbliži vektor u slici linearnog operatora A , tj. u potprostoru $R(A)$. Dakle, AX je ortogonalna projekcija od \mathcal{F} na spomenuti potprostor. Označimo li tu ortogonalnu projekciju sa \mathcal{F}' , traženi X bit će rješenje sustava

$$AX = \mathcal{F}'. \quad (1.6)$$

U nastavku ćemo koristiti sljedeći teorem (za dokaz vidi npr. [21]):

Teorem 1.4. *Za svaku matricu $A \in \mathbb{R}^{m \times n}$ vrijedi ortogonalna dekompozicija:¹*

$$N(A^\tau) \oplus R(A) = \mathbb{R}^m. \quad (1.7)$$

Sada pokažimo da je sustav (3.1.1) ekvivalentan sljedećem:

$$A^\tau AX = A^\tau \mathcal{F}'. \quad (1.8)$$

¹ $N(A^\tau)$ označava jezgru linearnog operatora A^τ .

Naime, množenjem (3.1.1) slijeva sa A^τ dobivamo (1.8). Preostaje pokazati da (1.8) \implies (3.1.1). Budući da je \mathcal{F}' definiran kao ortogonalna projekcija toka \mathcal{F} na potprostor $R(A)$, on je jednak AX' za neki $X' \in R^n$. Iz (1.8) sada slijedi

$$0 = A^\tau(AX - \mathcal{F}') = A^\tau(AX - AX') = A^\tau A(X - X'). \quad (1.9)$$

Dakle, vektor $A(X - X')$ nalazi se u $N(A^\tau)$. Ali on se očito nalazi i u $R(A)$. Budući da su (po Teoremu 1.4) ovi potprostori međusobno ortogonalni, u njihovom je presjeku samo nulvektor, pa slijedi $A(X - X') = 0$, odnosno $AX = AX' = \mathcal{F}'$, a to je (3.1.1).

Usredotočimo se sada na sustav (1.8), koji osim nepoznanice X sadrži i nepoznanicu \mathcal{F}' . Budući da je \mathcal{F}' dobiven kao ortogonalna projekcija toka \mathcal{F} na potprostor $R(A)$, slijedi da \mathcal{F} možemo zapisati kao $\mathcal{F}' + y$, pri čemu je y okomit na $R(A)$. No, po Teoremu 1.4 to znači da je $y \in N(A^\tau)$, tj. $A^\tau y = 0$. Sada množenjem jednakosti $\mathcal{F} = \mathcal{F}' + y$ slijeva sa A^τ slijedi $A^\tau \mathcal{F} = A^\tau \mathcal{F}'$, pa sustav (1.8) možemo zapisati kao

$$A^\tau AX = A^\tau \mathcal{F}. \quad (1.10)$$

Primijetimo da je $A^\tau A$ zapravo Laplaceova matrica L . Produkt $A^\tau \mathcal{F}$ označimo znakom ∇ i raspišimo njegove komponente, koristeći pritom definiciju matrice A :

$$\nabla_v = \sum_{\alpha=1}^m A_{v,\alpha}^\tau \mathcal{F}_\alpha = \sum_{\alpha \text{ ulazi u } v} \mathcal{F}_\alpha - \sum_{\alpha \text{ izlazi iz } v} \mathcal{F}_\alpha, \quad (1.11)$$

što znači da je ∇_v razlika između ukupnog toka koji ulazi u v i ukupnog toka koji izlazi iz v . Zato vektor ∇ zovemo *razlikom toka* (eng. flow-difference). Uočimo da iz (1.11) slijedi

$$\sum_{v=1}^n \nabla_v = 0. \quad (1.12)$$

Sustav (1.10) sada možemo pisati kao

$$LX = \nabla. \quad (1.13)$$

Primijetimo da ovaj sustav nema jedinstveno rješenje. Naime, ako je k broj povezanih komponenti grafa \mathcal{G} , poznato je da defekt matrice L iznosi k . Dakle, prostor rješenja sustava (čija je dimenzija jednaka defektu od L) najmanje je jednodimenzionalan, dakle beskonačan.

Drugi je način da uočimo beskonačan broj rješenja sljedeći: sustav (1.13), kao što smo pokazali, ekvivalentan je sustavu $AX = \mathcal{F}'$, a taj se svodi na jednadžbe oblika $X_u -$

$X_v = \mathcal{F}'_{(u,v)}$, koje očitó ostaju zadovoljene ako svim komponentama vektora X dodamo proizvoljan realan broj c . No odatle zaključujemo i da uvijek postoji rješenje X za koje vrijedi $\sum_{v=1}^n X_v = 0$, jer za bilo koje rješenje X' možemo svakoj njegovoj komponenti dodati $c = -(1/n) \sum_{v=1}^n X'_v$ i tako zadovoljiti spomenuti uvjet.

Uvjet $\sum_{v=1}^n X_v = 0$ smanjuje dimenziju prostora rješenja za 1 pa u slučaju povezanoga grafa ($k = 1$) uz njega dobivamo jedinstveno rješenje. Općenito je potrebno dimenziju prostora rješenja smanjiti za k , što činimo tako da zahtijevamo $\sum X_v = 0$ na svakoj povezanoj komponenti grafa \mathcal{G} . Na taj način svaka povezana komponenta ima jednaku ukupnu vrijednost potencijala (koja iznosi 0), što je i prirodno jer između njih ne postoje nikakvi bridovi, tj. preferencije.

Konačno možemo definirati potencijal (ne nužno konzistentnog) grafa preferencije.

Definicija 1.5. *Potencijal grafa preferencije rješenje je sustava jednadžbi*

$$LX = \nabla, \quad (1.14)$$

uz dodatni uvjet $\sum X_v = 0$ na svakoj povezanoj komponenti grafa, radi jedinstvenosti.

Iskažimo sada teorem koji smo dokazali prilikom izvoda sustava $LX = \nabla$.

Teorem 1.6. *Potencijal je najbolje približno rješenje sustava $AX = \mathcal{F}$ u smislu najmanjih kvadrata, tj. on minimizira udaljenost $\|AX - \mathcal{F}\|$.*

Primjer 1.7. *Potencijal grafa sa Slike 1.1 dobivamo rješavanjem sustava*

$$\begin{bmatrix} 2 & -1 & -1 & 0 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 2 & 0 \\ 0 & -1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} X = \begin{bmatrix} -3 \\ 9 \\ -4 \\ -2 \\ 0 \end{bmatrix} \implies X = \begin{bmatrix} -1.0833 \\ 2.25 \\ -1.4166 \\ 0.25 \end{bmatrix}.$$

Naime, prva četiri retka matrice na lijevoj strani sustava čine Laplaceovu matricu L , prve četiri komponente vektora na desnoj strani sustava čine ∇ , a peti redak sustava dodan je tako da ostvari uvjet $\sum_{v=1}^4 X_v = 0$. Provjerimo koliko dobro dobiveni potencijal odražava preferencije iz grafa:

$$X_b - X_a = 3.333 \approx 3 = \mathcal{F}_{(b,a)},$$

$$X_c - X_a = -0.333 \approx 0 = \mathcal{F}_{(c,a)},$$

$$X_b - X_c = 3.666 \approx 4 = \mathcal{F}_{(b,c)},$$

$$X_b - X_d = 2 = \mathcal{F}_{(b,d)},$$

čime možemo biti zadovoljni (sjetimo se da graf nije konzistentan).

1.3 Svojstva potencijala

Teoremi koji slijede autorov su doprinos te omogućuju da potencijal računamo iz sustava koji ima jedinstveno rješenje, bez potrebe za dodavanjem uvjeta o zbroju elemenata X_v .

Teorem 1.8. *Potencijal povezanoga grafa preferencije jedinstveno je rješenje odgovarajućeg sustava*

$$(L + E)X = \nabla, \quad (1.15)$$

pri čemu je E matrica sastavljena od samih jedinica.

Dokaz. Ako je X potencijal povezanoga grafa preferencije, onda je $LX = \nabla$ i $\sum_{v=1}^n X_v = 0$. Potonje povlači $EX = 0$ i sada lako slijedi (1.15).

Obrnuto, neka je X neko rješenje sustava (1.15). Koristeći (1.12) slijedi

$$0 = \sum_{i=1}^n \nabla_i = \sum_{i=1}^n \sum_{j=1}^n (L + E)_{ij} X_j = \sum_{j=1}^n X_j \left(\sum_{i=1}^n (L + E)_{ij} \right).$$

Iz definicije Laplaceove matrice lako se vidi da je izraz u zagradi jednak n , pa dobivamo $0 = n \sum_{j=1}^n X_j$. Odatle slijedi i $EX = 0$. Dakle, dobili smo

$$LX = \nabla - EX = \nabla,$$

$$\sum_{v=1}^n X_v = 0,$$

što znači da je X potencijal. □

Slijedi poopćenje prethodnog teorema.

Teorem 1.9. *Potencijal (ne nužno povezanoga) grafa preferencije jedinstveno je rješenje odgovarajućeg sustava*

$$(L + C)X = \nabla, \quad (1.16)$$

pri čemu je C matrica takva da je $C_{ij} = 1$ ako vrhovi i, j pripadaju istoj povezanoj komponenti grafa, a inače $C_{ij} = 0$.

Dokaz. Neka graf preferencije ima k povezanih komponenti. Laplaceova matrica L može se napisati kao zbroj k Laplaceovih matrica od kojih svaka opisuje samo jednu povezanu komponentu grafa: $L = L^{(1)} + \dots + L^{(k)}$. Isto vrijedi za vektor $\nabla = \nabla^{(1)} + \dots + \nabla^{(k)}$.

Neka je $E^{(i)}$ matrica takva da je $E_{uv}^{(i)} = 1$ ako vrhovi u, v oba pripadaju i -toj povezanoj komponenti grafa, a $E_{uv}^{(i)} = 0$ inače. Primijetimo da je $C = E^{(1)} + \dots + E^{(k)}$.

Iz prethodnog teorema slijedi da potencijal i -te povezane komponente dobivamo rješavanjem sustava $(L^{(i)} + E^{(i)})X = \nabla^{(i)}$ čije je rješenje jedinstveno (na indeksima koji pripadaju toj povezanoj komponenti). Budući da ovi sustavi za različite povezane komponente grafa nemaju zajedničkih ne-nul elemenata, možemo ih spojiti u ekvivalentan sustav koji će biti upravo $(L + C)X = \nabla$. Dakle, i taj sustav ima jedinstveno rješenje koje je sastavljeno od potencijala povezanih komponenata grafa, tj. radi se o potencijalu cijeloga grafa preferencije. \square

Opišimo sada nekoliko posebnih slučajeva u kojima potencijal računamo lakše nego rješavanjem sustava jednadzbi.

Ako je graf konzistentan (takvo je npr. bilo koje stablo), upisujemo npr. $X_1 = 0$ i potom se iz vrha 1 širimo bridovima grafa. Svaki put kad se širimo nekim bridom, potencijal barem jednog od vrhova tog brida već je upisan, a potencijal drugog vrha brida upisujemo tako da zadovoljimo jednakost $X_u - X_v = \mathcal{F}_{(u,v)}$ za trenutni brid. Budući da je graf konzistentan, nikad nećemo upisati vrijednost koja će pokvariti jednakost za neki drugi brid. Na koncu svim upisanim vrijednostima potencijala povezane komponente dodajemo odgovarajući c tako da za tu komponentu vrijedi $\sum X_v = 0$. Ovaj postupak ponovimo za svaku povezanu komponentu.

Ako je graf potpun, tada je njegov potencijal

$$X = \frac{1}{n}\nabla, \quad (1.17)$$

što se lako provjeri uvrštavanjem u $LX = \nabla$, uz korištenje jednakosti $\sum_{v=1}^n \nabla_v = 0$ i činjenice da Laplaceova matrica potpunoga grafa ima $n - 1$ na glavnoj dijagonali te -1 na ostalim mjestima.

Dokažimo sada svojstvo monotonosti potencijala [23], koje nam jamči da će se povećanjem intenziteta preferencije $u > v$ povećati i razlika između induciranih vrijednosti X_u i X_v , što bismo i očekivali.

Propozicija 1.10 (Monotonost). *Neka su \mathcal{F} i \mathcal{F}' tokovi preferencije grafa (V, \mathcal{A}) takvi da je $\mathcal{F}'_\alpha > \mathcal{F}_\alpha$ za brid $\alpha = (u, v) \in \mathcal{A}$, dok je $\mathcal{F}'_\beta = \mathcal{F}_\beta$ za sve ostale bridove $\beta \in \mathcal{A}$, $\beta \neq \alpha$. Ako je X potencijal toka \mathcal{F} , a X' potencijal toka \mathcal{F}' , vrijedi*

$$X'_u - X'_v > X_u - X_v.$$

Dokaz. Koristimo standardne oznake: A i L su matrica incidencije i Laplaceova matrica grafa (V, \mathcal{A}) , dok su ∇ , ∇' i $\nabla_{\mathcal{F}' - \mathcal{F}}$ redom razlike tokova \mathcal{F} , \mathcal{F}' i $\mathcal{F}' - \mathcal{F}$; drugim riječima, $\nabla = A^T \mathcal{F}$, $\nabla' = A^T \mathcal{F}'$ i $\nabla_{\mathcal{F}' - \mathcal{F}} = A^T (\mathcal{F}' - \mathcal{F})$.

Pretpostavimo da je \mathcal{G} povezan – ako nije, dovoljno je promatrati povezanu komponentu koja sadrži brid α . Neka je $\mathcal{F}'_\alpha - \mathcal{F}_\alpha = c > 0$, tj. $\mathcal{F}' - \mathcal{F} = c \cdot e_\alpha$. Zbog povezanosti, $AY = 0$ ako i samo ako je $Y = \text{const.}$ Pokazat ćemo da je $X' - X \neq \text{const.}$, što povlači $A(X' - X) \neq 0$, pa dalje imamo

$$\begin{aligned}
0 &< \langle A(X' - X), A(X' - X) \rangle \\
&= \langle A^\top A(X' - X), X' - X \rangle \\
&= \langle L(X' - X), X' - X \rangle \\
&= \langle \nabla' - \nabla, X' - X \rangle \\
&= \langle A^\top(\mathcal{F}' - \mathcal{F}), X' - X \rangle \\
&= \langle \mathcal{F}' - \mathcal{F}, A(X' - X) \rangle \\
&= \langle c \cdot e_\alpha, A(X' - X) \rangle \\
&= c(\langle e_\alpha, AX' \rangle - \langle e_\alpha, AX \rangle) \\
&= c((X'_u - X'_v) - (X_u - X_v)),
\end{aligned}$$

i propozicija slijedi. Preostaje dokazati da je $X' - X \neq \text{const.}$ Pretpostavimo suprotno: $X' - X = \text{const.}$ Budući da je $\sum X'_v = 0$ i $\sum X_v = 0$, nužno je $\text{const} = 0$, tj. $X' = X$. Ali tada je $\nabla = LX = LX' = \nabla'$, što je nemoguće jer je očito $\nabla \neq \nabla'$ u vrhovima u i v . \square

Za kraj se još jednom osvrnimo na važnost potencijala. Potencijal ne može u potpunosti opisati graf preferencije, ali budući da on zadovoljava jednakosti $X_u - X_v \approx \mathcal{F}_{(u,v)}$, komponente potencijala interpretiramo kao brojčane vrijednosti alternativa inducirane tokom preferencije \mathcal{F} . Te vrijednosti mogu se koristiti za stvaranje ljestvice poretka alternativa na temelju iskazanih preferencija (u je iznad v ako je $X_u > X_v$), iako potencijal nosi više informacija od pukog uređaja.

Poglavlje 2

Agregacija grafova preferencije

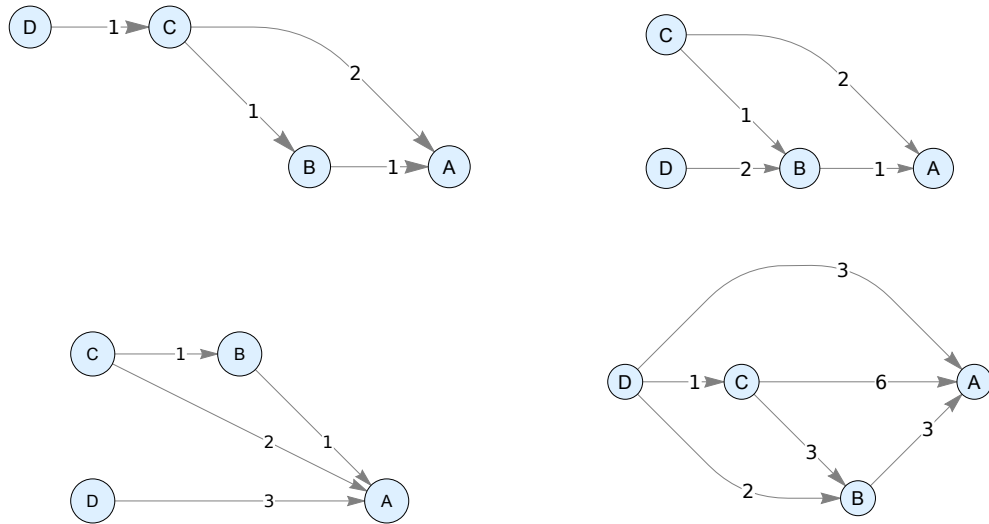
U prethodnom poglavlju vidjeli smo kako se iz pojedinoga grafa preferencije dobivaju odgovarajuće vrijednosti alternativa. Međutim, u situacijama kao što je grupno odlučivanje rezultate treba izvući iz više grafova preferencije – svaki graf odgovara jednom sudioniku. Također, u višekriterijskom odlučivanju svaki kriterij možemo smatrati jednim sudionikom koji daje svoj graf preferencije. Ponašanje čovjeka također je rezultat više ciljeva i interesa koji su nerijetko u međusobnom konfliktu [20]. U izbornim procedurama (o kojima će biti riječ u sljedećem poglavlju) glasači su sudionici u grupnom odlučivanju. U svakom od ovih slučajeva, preferencije različitih aktera koji često nisu međusobno suglasni treba na neki način agregirati.

Osim upravo spomenutih primjena te onih navedenih u Uvodu, agregacija preferencija javlja se u područjima kao što su pretraživanje informacija (engl. *information retrieval*) [3], teorija društvenog izbora [4], strojno učenje [18], dijagnostika [15], psihologija [13] i mnogim drugima.

2.1 Grupni (multi)graf preferencije

Odlučimo li se za graf preferencije kao osnovnu strukturu, ideja agregacije jest konstrukcija grupnoga grafa preferencije, onoga koji predstavlja preferencije cijele skupine aktera. U kontekstu grupnog odlučivanja, za dobivanje grupnoga grafa [22] predlaže zbrojiti tokove odgovarajućih bridova grafova preferencije. Sličnu metodu predlaže [7], ali za vrlo specifične grafove: potpune grafove kojima su težine (tokovi) bridova iz skupa $\{1, 1/2\}$. Agregacija zbrajanjem tokova prikladna je za potpune grafove. Ali kada grafovi nisu potpuni, ova metoda ne razlikuje brid koji ne postoji od brida koji ima tok 0 (jednaka preferencija alternativa). Slika 2.1 pokazuje primjer ove manjkavosti: tri osobe međusobno se potpuno slažu (njihovi grafovi preferencije imaju isti potencijal: $X_A = 1.5$, $X_B = 0.5$, $X_C = -0.5$,

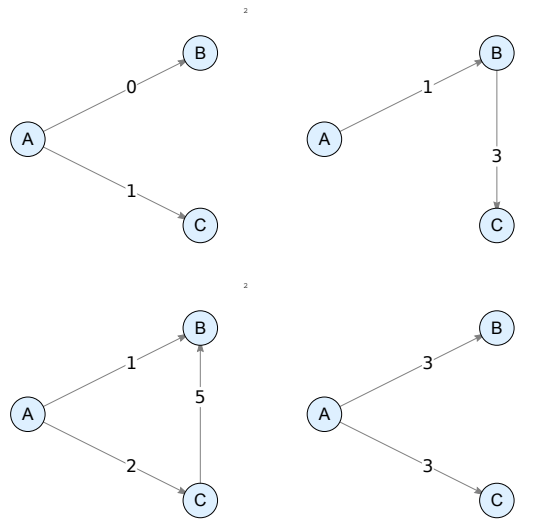
$X_D = -1.5$), ali grupni graf dobiven zbrajanjem tokova ne samo da ima drugačiji potencijal ($X_A = 3, X_B = 0.5, X_C = -2, X_D = -1.5$), nego taj potencijal povlači i drugačije rangiranje ($D > C$).



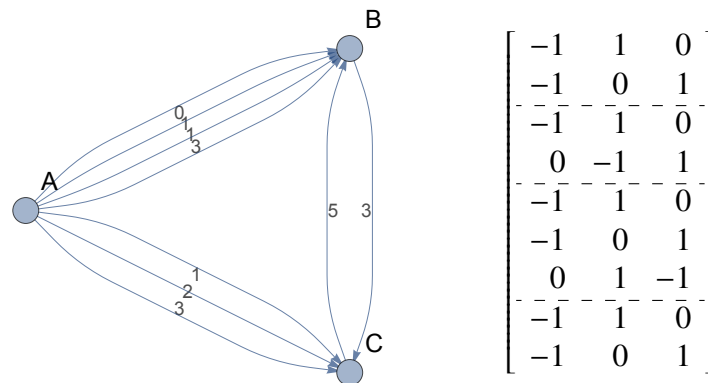
Slika 2.1: Agregacija predložena u [22]. Donji-desni graf dobiven je zbrajanjem tokova triju preostalih grafova.

Ovdje stoga predložemo drugi način agregacije. Umjesto da zbrajamo bridove grafova preferencije, mi ćemo ih sve ubaciti u grupni graf. Budući da ćemo tako dobiti višestruke bridove između istih dvaju alternativa, riječ je o multigrafu.

Preciznije, *preferencijski multigraf* $\mathcal{G} = (V, \mathcal{A}, \mathcal{F})$ usmjereni je multigraf dobiven spajanjem pojedinačnih grafova $\mathcal{G}_i = (V, \mathcal{A}_i, \mathcal{F}^{(i)})$, $i \in \{1, \dots, k\}$ koji sadrži sve bridove (s odgovarajućim tokovima) iz ovih grafova. Dakle, \mathcal{A} je multiset bridova dobiven spajanjem skupova \mathcal{A}_i – specijalno, $card(\mathcal{A}) = \sum_{i=1}^k card(\mathcal{A}_i)$. Paralelni i suprotni bridovi uobičajeni su i nastaju kada više osoba uspoređuje isti par alternativa. Budući da je svaki element od \mathcal{F} tok odgovarajućeg brida, vektor \mathcal{F} dobivamo vertikalnim nadovezivanjem vektora $\mathcal{F}^{(i)}$. Budući da svaki redak matrice incidencije opisuje jedan brid, matricu incidencije A multigrafa dobivamo vertikalnim nadovezivanjem matrica incidencije $A^{(i)}$ odgovarajućih grafova \mathcal{G}_i . Na primjer, multigraf na Slici 2.3 dobiven je spajanjem grafova na Slici 2.2.



Slika 2.2: Grafovi čijom agregacijom dobivamo multigraf na Slici 2.3.



Slika 2.3: Multigraf dobiven spajanjem grafova sa Slike 2.2. Desno je prikazana njegova matrica incidencije, sastavljena od matrica incidencije individualnih grafova.

2.2 Grupni potencijal

Budući da preferencijski multigraf sadrži sve preferencije iz grafova skupine, njegov potencijal dobar je predstavnik skupine i koristit ćemo ga za dobivanje rezultata, tj. vrijednosti alternativa. Da bismo taj grupni potencijal efikasno računali, dokazat ćemo najprije neka svojstva predložene agregacije.

Propozicija 2.1. Laplaceova matrica L multigrafa \mathcal{G} zbroj je Laplaceovih matrica $L^{(i)}$ odgovarajućih grafova \mathcal{G}_i .

Dokaz.

$$L = A^\tau A = \left[(A^{(1)})^\tau \dots (A^{(k)})^\tau \right] \begin{bmatrix} A^{(1)} \\ \dots \\ A^{(k)} \end{bmatrix} = \sum_{i=1}^k (A^{(i)})^\tau A^{(i)} = \sum_{i=1}^k L^{(i)}.$$

□

Propozicija 2.2. Razlika toka ∇ multigrafa \mathcal{G} zbroj je razlika toka $\nabla^{(i)}$ odgovarajućih grafova \mathcal{G}_i .

Dokaz.

$$\nabla = A^\tau \mathcal{F} = \left[(A^{(1)})^\tau \dots (A^{(k)})^\tau \right] \begin{bmatrix} \mathcal{F}^{(1)} \\ \dots \\ \mathcal{F}^{(k)} \end{bmatrix} = \sum_{i=1}^k (A^{(i)})^\tau \mathcal{F}^{(i)} = \sum_{i=1}^k \nabla^{(i)}.$$

□

Kao što smo definirali u Poglavlju 1, potencijal je rješenje sustava

$$LX = \nabla, \quad \sum X_v = 0. \quad (2.1)$$

Propozicije koje smo upravo dokazali omogućuju da izračunamo ovaj potencijal bez eksplisite konstrukcije multigrafa. Čim znamo Laplaceove matrice i razlike toka individualnih grafova \mathcal{G}_i , gornji sustav dobivamo izravno:

$$(L^{(1)} + \dots + L^{(k)})X = \nabla^{(1)} + \dots + \nabla^{(k)}. \quad (2.2)$$

Na prvi pogled, u gornjoj jednadžbi izgleda da svi individualni grafovi jednako pridonose grupnom potencijalu: naprosto zbrajamo Laplaceove matrice i razlike toka. Ali neke Laplaceove matrice gušće su od drugih i stoga više pridonose rezultatu. Na primjer, promotrimo agregaciju sljedećih dvaju grafova: potpunoga grafa \mathcal{G}_1 i rijetkoga grafa \mathcal{G}_2 (koji je npr. lanac ili stablo). Budući da \mathcal{G}_1 sadrži mnogo više bridova nego \mathcal{G}_2 , agregacijski multigraf, kao i njegova Laplaceova matrica, odražavat će \mathcal{G}_1 mnogo više nego \mathcal{G}_2 . To je ponekad poželjno: možda želimo da utjecajnost sudionika/glasača ovisi o količini informacija koje je iskazao. U tom je slučaju jednadžba (2.2) prikladna. Ali u drugim slučajevima akteri bi trebali rezultatima pridonositi jednako ili u skladu sa zadanim težinama $w_i > 0$.

Prethodni odlomak sugerira da u jednadžbi (2.2) doprinos grafa preferencije \mathcal{G}_i odgovara broju njegovih bridova $\text{card}(\mathcal{A}_i)$, ili ekvivalentno, "1-normi po elementima" Laplaceove matrice $L^{(i)}$, definiranoj kao

$$\|L\| := \|\text{vec}(L)\|_1 = \sum_{i,j} |L_{ij}|, \quad (2.3)$$

jer ova norma iznosi četiri puta broj bridova (brid (i, j) za 1 doprinosi $|L_{ij}|$, $|L_{ji}|$, $|L_{ii}|$ i $|L_{jj}|$).

Srećom, ovu normu možemo lako podesiti skaliranjem $L^{(i)}$ i $\nabla^{(i)}$ bilo kojim $\alpha_i > 0$. Skaliranje je opravdano time što ono ne mijenja potencijal odgovarajućega grafa preferencije \mathcal{G}_i :

$$L^{(i)}X^{(i)} = \nabla^{(i)} \iff (\alpha_i L^{(i)})X^{(i)} = \alpha_i \nabla^{(i)}. \quad (2.4)$$

Dakle, naš sustav bit će oblika

$$(\alpha_1 L^{(1)} + \dots + \alpha_k L^{(k)})X = \alpha_1 \nabla^{(1)} + \dots + \alpha_k \nabla^{(k)}, \quad (2.5)$$

te preostaje odabrati α_i ($i = 1, \dots, k$) tako da za sve $i, j \in \{1, \dots, k\}$ vrijedi

$$\frac{\|\alpha_i L^{(i)}\|}{\|\alpha_j L^{(j)}\|} = \frac{w_i}{w_j}, \quad (2.6)$$

pri čemu je w_i zadana težina sudionika/kriterija opisanoga grafom preferencije \mathcal{G}_i . Ovo dobivamo stavljanjem

$$\alpha_i = \frac{w_i}{\|L^{(i)}\|}, \quad i = 1, \dots, k. \quad (2.7)$$

Svime ovime opravdali smo sljedeću definiciju.

Definicija 2.3. Za zadane težine $w_1, \dots, w_k > 0$, težinski grupni potencijal grafova preferencije $\{\mathcal{G}_1, \dots, \mathcal{G}_k\}$ nad istim skupom alternativa V rješenje je sustava

$$\left(\sum_{i=1}^k \frac{w_i}{\|L^{(i)}\|} L^{(i)} \right) X = \sum_{i=1}^k \frac{w_i}{\|L^{(i)}\|} \nabla^{(i)}, \quad (2.8)$$

uz uvjet $\sum X_v = 0$ na svakoj povezanoj komponenti multigrafa, radi jedinstvenosti. Pritom su $L^{(i)}$ i $\nabla^{(i)}$ redom Laplaceova matrica i razlika toka grafa \mathcal{G}_i .

Ako su sve zadane težine $w_i = 1$, ispuštamo riječ "težinski", tj. X zovemo grupnim potencijalom.

Na koncu dokažimo jednostavno, ali važno svojstvo opisane agregacije.

Propozicija 2.4. *Ako svi grafovi skupine imaju jednake potencijale, neovisno o njihovoj individualnoj (ne)konzistentnosti, težinski grupni potencijal bit će jednak tom potencijalu.*

Dokaz. Neka je X potencijal svakog od grafova $\{\mathcal{G}_1, \dots, \mathcal{G}_k\}$. Budući da je

$$L^{(i)}X = \nabla^{(i)}, \quad \forall i \in \{1, \dots, k\},$$

onda očito vrijedi i

$$(\alpha_1 L^{(1)} + \dots + \alpha_k L^{(k)})X = \alpha_1 \nabla^{(1)} + \dots + \alpha_k \nabla^{(k)}.$$

Dakle, X je rješenje sustava za težinski grupni potencijal. Još valja pokazati da je $\sum X_v = 0$ na svakoj povezanoj komponenti multigrafa. Zbog načina njegove konstrukcije, za bilo koju povezanu komponentu multigrafa primijetimo da se ona sastoji od jedne ili više povezanih komponenta grafa \mathcal{G}_1 (ili bilo kojeg \mathcal{G}_i). Budući da za svaku od tih komponenta grafa \mathcal{G}_1 vrijedi $\sum X_v = 0$, to vrijedi i za njihovu uniju. \square

Poglavlje 3

Univerzalna izborna procedura

U ovom poglavlju opisujemo primjenu agregacije grafova preferencije na izbornu proceduru. Svaka od postojećih izbornih procedura ovisi o određenom obliku glasačkog listića. U radu [23], L. Čaklović i autor stoga predlažu općenitu izbornu proceduru koja ne ovisi o obliku glasačkog listića, tj. primjenjiva je za bilo koji njegov oblik.

Njome rješavamo i sljedeći problem: svaki oblik glasačkog listića ima svoje nedostatke. Na primjer, široko korišteni listić je onaj na kojemu glasač zaokružuje samo jednu (najdražu) alternativu¹ i na taj način ne daje mnogo informacija o svojim preferencijama: ne može iskazati koje alternative mu se sviđaju osim najdraže, a što je još važnije, ne može izraziti nezadovoljstvo, tj. iskazati koje alternative ne voli. Bez obzira na to, ovakav listić najčešće se koristi u izbornoj proceduri. Razlog je njegova jednostavnost. Precizniji glasački listići, kao što su rangiranje svih alternativa od najdraže do najmrže, dodjela bodova ili usporedba u parovima, suviše su složeni za mnoge glasače, pogotovo kad je broj alternativa velik. Povećanjem preciznosti gubimo na jednostavnosti i obratno. Nema glasačkog listića koji je vrlo precizan i široko primjenjiv.

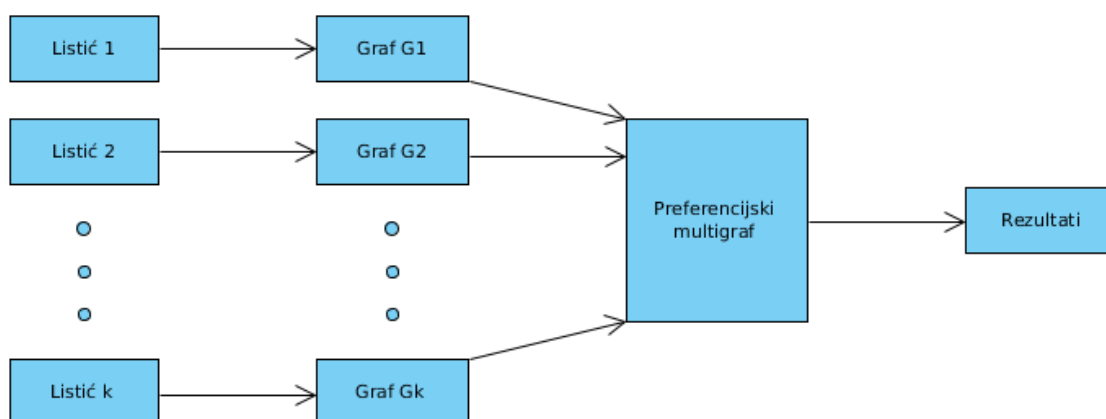
Univerzalna izborna procedura rješava ovaj problem time što ne zahtijeva isti glasački listić za sve glasače. Neki glasači žele jednostavan listić, drugi su spremni potrošiti više vremena da precizno iskažu svoje preferencije, a mnogi su između. Što ako svaki glasač dobije listić koji mu odgovara? Može li u tom slučaju utjecaj (težina) na rezultate biti jednak za sve glasače, neovisno o tome koliko su informacija o svojim preferencijama iskazali na svojim listićima? Štoviše, mogu li se težine proizvoljno postaviti? Univerzalna izborna procedura podržava sve ove ideje. Dakle, ne samo da podržava bilo koji izabrani oblik glasačkog listića, nego i različite oblike listića u istim izborima.

Istaknimo da je ova metoda primjenjiva i na grupno odlučivanje općenito, kao i na

¹U izbornim procedurama, *alternative* su najčešće kandidati koji se natječu na izborima.

višekriterijsko odlučivanje: svaki kriterij daje svoj "listić" (ponovno, mogući su različiti oblici listića), a procedura koju ćemo opisati potom daje ukupne vrijednosti alternativa s obzirom na ove kriterije i njihove važnosti (težine). Stoga se u nastavku poglavlja riječ *glasač* može zamijeniti i riječju *kriterij*.

Procedura se sastoji od dviju faza: najprije se glasački listići pretvaraju u grafove preferencije, a potom se dobiveni grafovi preferencije agregiraju, čime dobivamo rezultate, tj. vrijednosti alternativa inducirane preferencijama glasača. Skica procedure dana je na Slici 3.1.



Slika 3.1: Skica univerzalne izborne procedure.

Agregacija grafova preferencije (druga faza univerzalne izborne procedure) već je opisana u prethodnome poglavlju pa stoga u nastavku opisujemo prvu fazu: konstrukciju tih grafova.

3.1 Pretvorba glasačkih listića u grafove preferencije

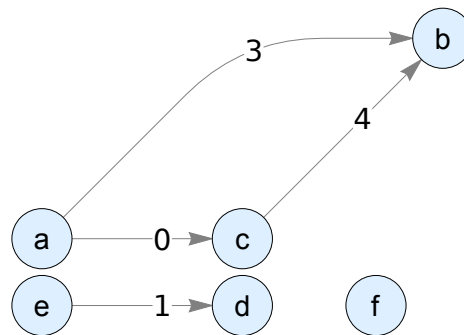
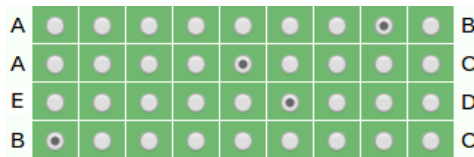
Preferencije glasača mogu biti dane u raznim oblicima; mnoge od njih opisat ćemo u ovoj sekciji. Pokazat ćemo kako se svi ti oblici iskazanih preferencija osobe mogu pretvoriti u graf preferencije, što omogućuje agregaciju. U primjerima koji slijede, preferencije su izražene između šest alternativa: a , b , c , d , e i f na razne načine i potom pretvorene u graf preferencije. Budući da se pretvorba često može učiniti na više načina, ove prijedloge ne treba shvatiti odviše strogo.

3.1.1 Usporedba parova

Ovo je standardni način konstrukcije grafa preferencije, oprimjeren i u Poglavlju 1. Pretpostavimo da glasač uspoređuje neke parove alternativa, bilo na način da samo navede koju od ponuđenih dviju alternativa preferira (prvu, drugu ili obje jednako), bilo tako da iskaže i intenzitet preferencije na brojčanoj ili verbalnoj skali.

Svaki od ovih načina može se svesti na brojčanu skalu. Naime, ako se navodi samo smjer preferencije, prirodno je intenzitet proglasiti jednakim 1. Ako pak glasač navodi intenzitet preferencije na verbalnoj skali, ona se lako prevodi u brojčanu (npr. jednaka preferencija - 0, slaba - 1, jaka - 2, apsolutna - 3).

Glasač ne mora nužno usporediti sve parove alternativa, čime bismo dobili potpun graf. Bilo koji skup uspoređenih parova dat će graf preferencije. Taj skup ne mora sadržavati sve alternative: podaci koji nedostaju ne predstavljaju problem ovoj metodi; alternativa koja nije ni sa čim uspoređena naprosto će biti predstavljena vrhom bez incidentnih bridova u grafu. Također, glasač ne mora biti konzistentan u svojim usporedbama. To je vrlo poželjno svojstvo jer zahtjev za konzistentnošću nije realan ako glasač navodi intenzitete preferencije ili ako vrši veći broj usporedbi.



Slika 3.2: Graf preferencije dobiven usporedbom parova.

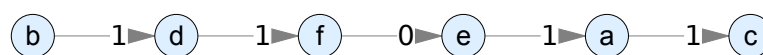
Kao primjer promotrimo glasački listić na Slici 3.2 (gornji dio). Usporedba dviju alternativa sastoji se od odabira jedne od devet točaka na skali; odabir sasvim lijeve točke znači da glasač snažno preferira lijevu alternativu nad desnom, i obrnuto; odabir točke na sredini označava jednaku preferenciju alternativa. U prvoj usporedbi, *b* je opisan kao bolji od *a*

za 3 točke, brojeći od središnje (neutralne) točke, što u grafu preferencije (Slika 3.2, donji dio) daje brid iz a u b s tokom 3. Ostali bridovi dobiveni su analogno.

Ako drugi glasači koriste drugačije skale, moguće je ujednačiti skale jednostavnim množenjem. U našem primjeru skala je $[-4, 4]$; ona se može pretvoriti npr. u skalu $[-5, 5]$ tako da se sve vrijednosti pomnože s 1.25.

3.1.2 Rangiranje svih alternativa

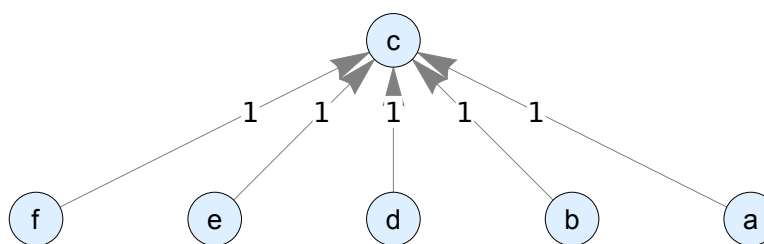
Glasač čiji je poredak $c > a > e \sim f > d > b$ prirodno se opisuje grafom preferencije na Slici 3.3.



Slika 3.3: Graf preferencije dobiven iz poretka svih alternativa.

3.1.3 Odabir najdraže alternative

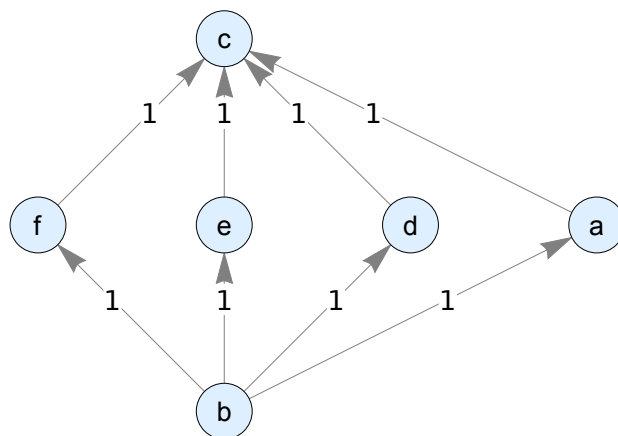
Pretpostavimo da je glasač naveo c kao najdražu alternativu među a, b, c, d, e, f . Budući da takav iskaz implicira usporedbe $c > a, c > b, c > d, c > e$ i $c > f$, dobivamo graf preferencije na Slici 3.4.



Slika 3.4: Graf preferencije nastao odabirom najdraže alternative.

3.1.4 Odabir najdraže i najmrže alternative

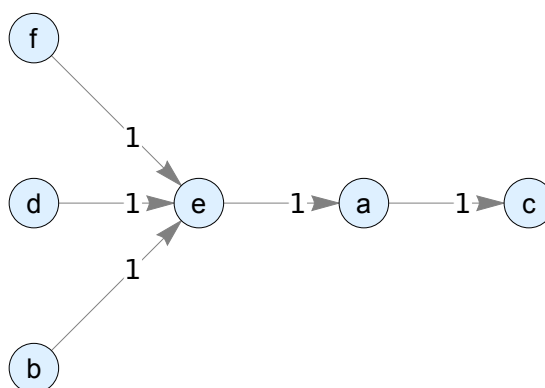
Pretpostavimo da je glasač naveo c kao najdražu i b kao najmržu alternativu. Budući da takav iskaz implicira usporedbe $c > a > b, c > d > b, c > e > b$ i $c > f > b$, dobivamo graf preferencije na Slici 3.5.



Slika 3.5: Graf preferencije nastao odabirom najdraže i najmrže alternative.

3.1.5 Rangiranje najdražih k alternativa

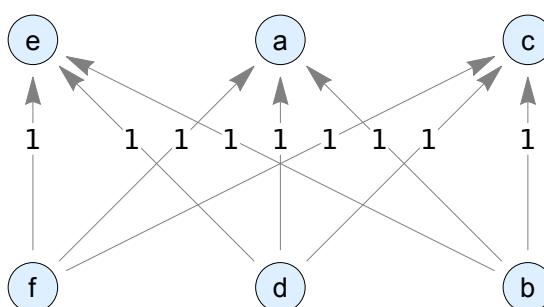
Pretpostavimo da je glasač kao najdraže 3 alternative naveo $c > a > e$. Implicirane usporedbe su $e > b$, $e > d$ i $e > f$, pa dobivamo graf preferencije na Slici 3.6.



Slika 3.6: Graf preferencije nastao rangiranjem najdražih k alternativa.

3.1.6 Navođenje najdražih k alternativa bez rangiranja

Pretpostavimo da je glasač kao najdraže 3 alternative naveo skup $\{a, c, e\}$, bez preciziranja poretka tih triju alternativa. Glasač implicira da mu je svaka od tih alternativa draža od svake od preostalih alternativa, što daje graf preferencije na Slici 3.7.



Slika 3.7: Graf preferencije nastao navođenjem najdražih k alternativa bez rangiranja.

3.1.7 Ocjenjivanje alternativa

Pretpostavimo da glasač ocjenjuje alternative ocjenama s fiksne skale, npr. od 1 do 10. *Razlike* ocjena koristimo kao tokove odgovarajućih bridova grafa.

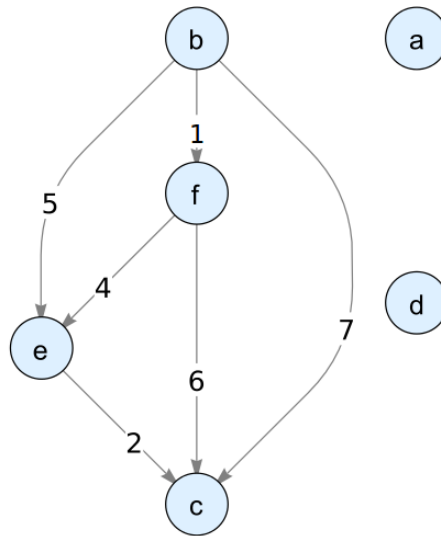
Ocjene koje nedostaju ne predstavljaju problem ovoj metodi; njihovo postojanje samo znači da dobiveni graf preferencije nije potpun. Ovo je vrlo poželjno svojstvo jer je ocjenjivanje svih alternativa prilično nerealan zahtjev kad je njihov broj velik ili kada glasač nema informacije o svim alternativama. Glasač može sam izabrati koje alternative će ocijeniti.

Kao primjer, pretpostavimo da je glasač ocijenio b, c, e i f sljedećim ocjenama sa skale $[1, 10]$: 2, 9, 7, 3, redom. Slika 3.8 prikazuje dobiveni graf preferencije. Ocjena alternative e veća je za 5 od ocjene alternative b , što daje brid iz b u e s tokom 5.

Skala ne mora biti brojčana: mnogima je lakše ocjenjivati na verbalnoj skali (npr. *izvrstan, dobar, osrednji, loš*). U ovom slučaju, verbalne ocjene moraju se pretvoriti u brojeve. Kao što je spomenuto u Odjeljku 3.1.1, lako je ujednačiti različite brojčane skale.

3.1.8 Dodjela bodova

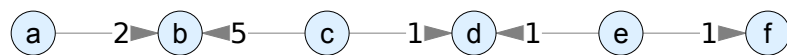
Pretpostavimo da glasač na raspolaganju ima određeni broj bodova – na primjer 10 – koje dodjeljuje alternativama tako da jače preferirane alternative dobiju više bodova. Pretposta-



Slika 3.8: Graf preferencije nastao ocjenjivanjem alternativa.

vimo da a, b, c, d, e, f dobiju 3, 5, 0, 1, 0, 1 bodova, redom.

Kao u prethodnome primjeru, razlike u broju bodova postaju tokovi odgovarajućih bridova grafa preferencije, što bi u ovom slučaju rezultiralo potpunim grafom jer je poznat broj bodova koje je dobila svaka alternativa. Ali u ovom potpunom grafu mnogi su bridovi redundantni i mogu se ukloniti. Dovoljno je umjesto potpunoga grafa načiniti lanac bridova koji prolazi svim vrhovima ($a - b - c - d - e - f$) tako da su tokovi bridova razlike odgovarajućih bodova (Slika 3.9), jer su time implicirane i sve ostale razlike koje nisu predstavljene bridovima. Takav graf ima isti potencijal kao i spomenuti potpuni graf. Kao i u prethodnim slučajevima, različite skale mogu se ujednačiti.

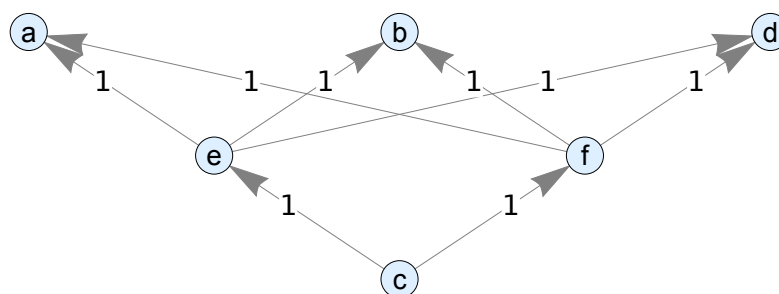


Slika 3.9: Graf preferencije nastao dodjelom bodova (3, 5, 0, 1, 0, 1).

3.1.9 Komentari

Principi iz gornjih primjera široko su primjenjivi: navedenim oblicima glasačkih listića nisu, naravno, iscrpljene sve mogućnosti. Na primjer, za izražavanje nezadovoljstva, glasač

na raspolaganju osim pozitivnih bodova može imati i određeni broj negativnih bodova koje dodjeljuje alternativama. Također, navedene oblike moguće je međusobno kombinirati u nove oblike. Kao ilustraciju promotrimo sljedeći iskaz: "najviše volim a , b i d , a najmanje c ", koji kombinira navođenje najdražih k alternativa i navođenje najmrže alternative. On implicira usporedbe $a > e$, $b > e$, $d > e$, $a > f$, $b > f$, $d > f$, $e > c$ i $f > c$, pa dobivamo graf preferencije na Slici 3.10. Takvi i mnogi drugi oblici lako se pretvaraju u grafove preferencije slijedeći principe iz gornjih primjera.



Slika 3.10: "Najviše volim a , b i d , a najmanje c ."

Kao što smo u jednom slučaju primijetili, neki od predloženih grafova mogu se pretvoriti u nešto drugačije, ekvivalentne grafove (u smislu da imaju isti potencijal): redundantni bridovi mogu se ukloniti. Zbog velikog broja mogućnosti nema smisla definirati stroga pravila za konstrukciju grafa preferencije.

Vidjeli smo da su mnogi od opisanih oblika fleksibilni za glasača u smislu da svaki od njih može izabrati *koliko* informacija će dati i *koje* će to informacije biti, što je iznimno poželjno svojstvo ove univerzalne izborne procedure.

3.2 Eksperiment

Opisanu metodu za ilustraciju smo primijenili na podatke s glasanja na Eurosong natjecanju 2005. godine; podaci su preuzeti iz rada [19] gdje su ih autori koristili za evaluaciju svog algoritma rangiranja. Na finalnom natjecanju Eurosonga 2005. sudjelovale su 24 pjesme iz isto toliko zemalja, a o poretku pjesama glasalo je 39 zemalja (osim zemalja koje su sudjelovale, glasale su i zemlje koje se nisu plasirale u finalno natjecanje). Tablica 3.1 prikazuje rezultate glasanja: svaki redak navodi bodove koje je odgovarajuća zemlja dodijelila pjesmama. Bodovi pojedine zemlje dodjeljivani su na sljedeći način: pjesma s

	GR	MA	RO	IS	LA	MD	SM	SZ	DE	NO	CR	HU	TU	BH	RU	AL	FY	CY	SW	UA	SP	UK	FR	GE
AL	12	4	5	3	0	0	6	0	0	0	2	0	8	0	0	10	7	0	0	0	0	1	0	
AN	4	0	7	8	10	0	0	1	3	2	0	6	0	0	0	0	0	0	0	12	0	5	0	
AU	4	5	6	1	0	2	12	0	0	0	8	0	7	10	0	3	0	0	0	0	0	0	0	
BE	0	5	1	8	6	7	3	10	0	4	0	2	0	0	12	0	0	0	0	0	0	0	0	
BG	12	8	7	6	5	1	0	0	4	3	0	2	10	0	0	0	0	0	0	0	0	0	0	
BH	6	0	2	0	0	4	10	0	0	3	12	1	8	0	5	7	0	0	0	0	0	0	0	
BU	12	0	8	0	0	6	4	0	0	1	2	5	3	0	0	0	7	10	0	0	0	0	0	
CR	5	4	0	0	7	1	12	3	0	0	6	0	10	0	2	8	0	0	0	0	0	0	0	
CY	12	6	8	0	1	2	10	4	0	3	0	7	0	0	0	0	0	0	0	5	0	0	0	
DE	2	10	3	5	6	0	0	1	12	0	0	8	4	0	0	0	0	7	0	0	0	0	0	
ES	0	4	0	1	10	6	0	12	5	8	2	3	0	0	7	0	0	0	0	0	0	0	0	
FY	7	0	2	0	0	5	10	0	0	0	8	1	4	3	0	12	0	6	0	0	0	0	0	
FI	3	8	0	5	4	0	0	10	2	12	1	0	0	0	7	0	0	0	6	0	0	0	0	
FR	8	7	5	10	0	2	6	0	0	0	0	3	12	0	0	1	0	0	0	4	0	0	0	
GE	12	8	0	5	7	1	3	4	6	0	2	0	10	0	0	0	0	0	0	0	0	0	0	
GR		8	5	0	1	7	6	3	0	4	0	2	0	0	0	10	0	12	0	0	0	0	0	
HU	12	5	10	8	1	4	2	3	6	0	7	0	0	0	0	0	0	0	0	0	0	0	0	
IC	2	4	5	0	3	8	0	7	10	12	1	6	0	0	0	0	0	0	0	0	0	0	0	
IR	2	10	5	6	12	0	0	3	7	4	0	0	0	1	0	0	0	0	0	0	8	0	0	
IS	7	10	12		6	4	0	2	3	1	0	8	0	0	0	0	0	0	0	0	0	5	0	
LA	0	5	0	2		8	1	12	4	6	7	3	0	0	10	0	0	0	0	0	0	0	0	
LI	1	2	0	3	12	10	0	8	4	5	6	0	0	0	7	0	0	0	0	0	0	0	0	
MA	6		7	8	10	2	0	1	3	5	0	0	0	0	0	0	12	0	0	0	4	0	0	
MD	4	0	7	6	12		1	0	0	3	0	0	0	0	10	0	0	0	5	8	0	0	2	
MO	0	5	4	12	0	0	6	8	10	0	7	0	0	0	0	0	1	0	3	0	0	0	2	
NE	10	5	3	7	0	0	4	0	8	1	2	0	12	6	0	0	0	0	0	0	0	0	0	
NO	4	10	6	5	8	0	0	3	12		2	0	0	7	0	0	0	0	1	0	0	0	0	
PL	1	0	7	0	4	3	0	6	5	8	2	10	0	0	0	0	0	0	12	0	0	0	0	
PO	3	0	12	5	6	10	0	4	1	0	0	2	0	0	0	0	0	0	0	7	8	0	0	
RO	10	2		7	0	12	6	0	4	0	5	8	3	0	0	0	0	1	0	0	0	0	0	
RU	4	12	0	8	5	10	6	7	0	0	1	3	0	0		0	0	0	2	0	0	0	0	
SM	12	0	3	0	0	5		0	0	2	10	6	0	4	0	8	7	1	0	0	0	0	0	
SL	2	1	0	0	7	3	10	6	0	4	12	0	0	8	0	0	5	0	0	0	0	0	0	
SP	8	7	12	6	0	4	0	0	10	3	0	5	0	0	0	0	0	0	2	1		0	0	
SW	12	6	2	1	3	0	4	5	10	8	0	0	0	7	0	0	0		0	0	0	0	0	
SZ	7	3	0	1	0	0	12		0	0	8	0	6	5	0	10	2	0	0	0	4	0	0	
TU	12	8	4	3	0	7	0	0	0	0	0	6		10	0	2	5	0	0	0	1	0	0	
UA	0	10	0	7	1	12	3	5	0	6	8	2	0	0	4	0	0	0		0	0	0	0	
UK	12	10	0	7	6	2	0	0	8	5	0	0	1	4	0	0	0	3	0	0	0	0	0	

Tablica 3.1: Rezultati glasanja s Eurosonga 2005. [19]: svaki redak opisuje glasačke preferencije odgovarajuće zemlje.

najviše glasova dobila je 12 bodova, sljedeća pjesma 10 bodova, sljedeća 8, pa 7, 6, 5, 4, 3, 2 i 1 bod, a preostale pjesme dobile su 0 bodova od dotične zemlje.

Kao u Odjeljku 3.1.8 (Dodjela bodova), svaku zemlju prikazali smo grafom preferencije, tj. lancem. Ti su grafovi potom agregirani metodom iz prethodnoga poglavlja, uz pomoć softvera *pyPot* autora L. Čaklovića. Rezultati su prikazani u Tablici 3.2.

Zanimljivo je da je poredak zemalja u stvarnim rezultatima Eurosonga 2005. (dobive-

Zemlja	Vrijednost potencijala
Gruzija	3.538
Malta	2.555
Rumunjska	1.684
Izrael	1.577
Latvija	1.570
Moldavija	1.414
Srbija i Crna Gora	1.110
Švicarska	0.918
Norveška	0.828
Danska	0.822
Hrvatska	0.545
Mađarska	0.071
Turska	-0.063
Bosna i Hercegovina	-0.408
Rusija	-0.968
Albanija	-1.106
Makedonija	-1.120
Cipar	-1.279
Ukrajina	-1.684
Švedska	-1.695
Španjolska	-1.734
Ujedinjeno kraljevstvo	-2.007
Francuska	-2.191
Njemačka	-2.376

Tablica 3.2: Rezultati opisane univerzalne izborne procedure za glasačke preferencije s Eurosonga 2005.

nima zbrajanjem bodova) jednak poretku dobivenom našom metodom (Tablica 3.2). Jedina je razlika što u stvarnome poretku postoji neriješen rezultat između Norveške i Danske te između Ukrajine i Švedske, dok u našem poretku nema neriješenih rezultata, tj. postoji malena razlika u potencijalu spomenutih zemalja. Zaključujemo da eksperiment potvrđuje opisanu univerzalnu izbornu proceduru.

Ako bismo metodu proveli nešto drugačije, na način da u pojedinom retku Tablice 3.1 umjesto razlike broja bodova uzmemo u obzir samo predznak te razlike i tako dobijemo grafove preferencije s bridovima težine 1 ili -1 (tzv. unimodularni tok), dobivena ljestvica poretka bila bi malo drugačija: GR, MA, IS, RO, MD, LA, NO, SZ, SM, DE, CR, HU, TU, BH, FY, AL, RU, CY, SW, UA, SP, UK, FR, GE. Takav pristup primjenjiv je kada nema dodjele bodova.

Poglavlje 4

Klasterizacija grafova preferencije

Klasterizacija ili klasterijska analiza općenit je zadatak grupiranja objekata na način da međusobno slični objekti pripadaju istoj grupi (klasteru). Ovisno o tipu objekata, vrsti i broju klastera koje tražimo, definiciji sličnosti (ili udaljenosti) objekata ili klastera, te o definiciji dobroga grupiranja, postoje različiti pristupi rješavanju ovog problema. Klasterizacija je jedna od ključnih tema rudarenja podataka (engl. *data mining*) te ima brojne primjene [16]. Koristi se u statističkoj obradi podataka, strojnom učenju, prepoznavanju uzoraka, analizi slika, bioinformatiki i biologiji [1]. U medicini, klasterizacija bolesti, lijekova ili simptoma pomaže u taksonomiji, dijagnostici i odabiru terapije [10]. U društvenim znanostima poput ekonomije, sociologije i psihologije, klasterizacija se često koristi za svrstavanje ispitanika u kategorije [2]. Također se koristi u analizi društvenih skupina [11].

Objekti koje svrstavamo u klaster najčešće se predstavljaju kao uređene k -torke. Na primjer, ako svakom objektu mjerimo k varijabli (osobina), onda izmjerene vrijednosti tih varijabli za pojedini objekt čine njegov vektor (x_1, \dots, x_k) . Ako su sve varijable brojčane (realni brojevi), kao što su težina, visina, kvocijent inteligencije i slično, tada nije teško definirati udaljenost objekata, odnosno klastera. Nakon eventualnog skaliranja vrijednosti varijabli, udaljenost dvaju objekata najčešće se računa kao norma razlike odgovarajućih vektora, pri čemu norma može biti 2-norma (što daje euklidsku udaljenost), 1-norma (što daje tzv. Manhattan udaljenost), max-norma (što daje Čebiševljevu udaljenost) ili neka p -norma (što daje udaljenost Minkowskog). Ako se odlučimo za sličnost umjesto udaljenosti, možemo je računati npr. kao kosinus kuta između dvaju vektora ili kao Pearsonov koeficijent sličnosti. Moguće je definirati sličnost/udaljenost i u slučaju da su vrijednosti varijabli kategorijske (npr. boja), a ne brojčane [9]. Udaljenost dvaju klastera moguće je definirati na više načina, npr. kao najmanju, najveću ili prosječnu udaljenost između odgovarajućih elemenata klastera, ili kao udaljenost dvaju odgovarajućih središta klastera. Središte klastera može se definirati npr. kao aritmetička sredina njegovih elemenata.

4.1 Klasterizacija skupine na temelju preferencija njezinih članova

Standardni algoritmi za klasterisku analizu nisu primjenjivi u slučajevima kad se objekt klasterizacije ne može jednostavno opisati uređenom k -torkom varijabli. Ako je riječ o osobama koje izražavaju svoje preferencije, realno je pretpostaviti da su preferencije dane usporedbama (o tome je bilo riječi u Uvodu). Ako dakle osoba izražava relativne preferencije, poput "više volim A nego B " ili "najviše volim A , B i D , a najmanje C ", onda nema jasnoga skupa osobina koje bismo za svakog ispitanika izmjerili, tj. nema jasnoga niza vrijednosti koji bi opisao pojedinog ispitanika i bio objekt pogodan za klasterizaciju. To pogotovo vrijedi kada su iskazane preferencije nepotpune u smislu da ne daju informacije o svim alternativama, što je realna pretpostavka ako je broj alternativa velik.

Tim problemom bavi se ovo poglavlje, koje predstavlja autorov doprinos. U njemu se pretpostavlja postojanje skupa alternativa nad kojim svaka osoba iz promatrane skupine izražava svoje preferencije, te se opisuje klasterizacija osoba na temelju njihovih preferencija: osobe sličnih preferencija želimo grupirati zajedno. Kao i za univerzalnu izbornu proceduru, važno svojstvo predložene metode njezina je općenitost: metoda ne zahtijeva da preferencije osoba budu iskazane u nekom određenom, propisanom obliku, nego funkcionira za iskaze raznih oblika, potpune i nepotpune, konzistentne i nekonzistentne. Neki oblici iskaza spomenuti su u prethodnom odlomku, a mnoge druge vidjeli smo kao glasačke listiće u prethodnome poglavlju. Ovdje ćemo umjesto *glasačkog listića* koristiti drugi naziv: skup iskaza pojedine osobe o svojim preferencijama zvat ćemo *preferencijskim upitnikom*.

Kako god izgledao preferencijski upitnik, on se može pretvoriti u graf preferencije, kao što je opisano u prethodnome poglavlju za glasačke listiće. Jedan graf predstavlja jednu osobu i temeljni je objekt predložene klasterizacije (za razliku od uređenih k -torki koje se koriste u drugim slučajevima). Dakle, prva je faza algoritma konstrukcija grafova preferencije iz preferencijskih upitnika, a druga je faza sama klasterizacija, tj. grupiranje dobivenih grafova u klaster tako da oni grafovi koji predstavljaju osobe sličnih preferencija budu grupirani zajedno. Ovisno o svrsi, grupiranje može biti *hijerarhijsko* (klasteri sadrže pod-klaster) i *particijsko* (daje disjunktne klaster). Prvim se dobiva grupna hijerarhija, a drugim se ulazni skup dijeli (particionira) na određeni broj grupa.

Neka je m broj objekata koje grupiramo – u našem slučaju to su grafovi preferencije $\mathcal{G}_1, \dots, \mathcal{G}_m$, pri čemu je $\mathcal{G}_i = (V, \mathcal{A}_i, \mathcal{F}^{(i)})$. Aglomerativna hijerarhijska klasterizacija [16] kreće od m klastera: u svakom je klasteru jedan objekt. U svakoj iteraciji algoritma, najbliža dva klastera spajaju se u novi, viši klaster, sve dok sve objekte ne spojimo u jedan klaster. Na taj način dobivamo hijerarhijsku strukturu klastera: na najvišem nivou samo je jedan klaster, u njemu se nalaze dva klastera, svaki od tih klastera dalje je podijeljen

na dva klastera, i tako dalje sve do pojedinačnih objekata. Hijerarhijska struktura prikazuje se stablom, tzv. dendrogramom. Da bismo ovaj algoritam iskoristili u našem slučaju, potrebno je definirati udaljenost dvaju klastera kako bismo u pojedinoj iteraciji algoritma mogli pronaći dva najbliža klastera koje treba spojiti.¹

Kako definirati udaljenost dvaju klastera koji sadrže grafove preferencije? U skladu s razmatranjima iz Poglavlja 2, klaster ćemo predstaviti njegovim grupnim potencijalom.

Definicija 4.1. *Neka su $P = \{\mathcal{G}_1, \dots, \mathcal{G}_k\}$ i $Q = \{\mathcal{H}_1, \dots, \mathcal{H}_l\}$ dva skupa (klastera) grafova preferencije nad istim skupom alternativa V . Ako je X grupni potencijal klastera P , a Y grupni potencijal klastera Q , udaljenost klastera P i Q definiramo kao udaljenost vektora X i Y :*

$$d(P, Q) := \|X - Y\|. \quad (4.1)$$

Dodatno opravdanje ove definicije sljedeća je njezina posljedica koja daje očekivanu formulu za udaljenost dvaju grafova preferencije.

Korolar 4.2. *Udaljenost singletona $\{\mathcal{G}\}$ i $\{\mathcal{H}\}$ iznosi $\|X - Y\|$, pri čemu je X potencijal grafa preferencije \mathcal{G} , a Y potencijal grafa preferencije \mathcal{H} .*

Dokaz. Iz Propozicije 2.4 slijedi da je grupni potencijal klastera $\{\mathcal{G}\}$ jednak X , a grupni potencijal klastera $\{\mathcal{H}\}$ jednak Y i tvrdnja sada slijedi iz definicije udaljenosti klastera. \square

4.2 Hijerarhijska klasterizacija

Sada možemo primijeniti aglomerativnu hijerarhijsku klasterizaciju koristeći udaljenost klastera koju smo definirali maloprije. Algoritam je bio ukratko opisan u prethodnom odjeljku, a sada ga izričemo nešto preciznije.

Pretpostavimo da smo na ulazu dobili grafove preferencije $\mathcal{G}_1, \dots, \mathcal{G}_m$. Algoritam cijelo vrijeme održava skup trenutnih klastera C . Na samom početku postoji m klastera: svaki graf preferencije \mathcal{G}_i nalazi se u svom klasteru $\{\mathcal{G}_i\}$. Algoritam vrši $m - 1$ iteracija, u svakoj spajajući dva najbliža klastera (s obzirom na udaljenost iz Definicije 4.1) u novi klaster, smanjujući tako broj trenutnih klastera za 1. Naravno, prilikom svakog spajanja treba spremati informaciju o klasterima koje smo spojili, jer jedino tako dobivamo izlaznu hijerarhijsku strukturu klastera.

Slijedi pseudokod algoritma.

¹Osim aglomerativne, postoji i razdvajajuća hijerarhijska klasterizacija, koja kreće od najvećeg klastera i rekurzivno ga dijeli na manje klastere. Ona se, međutim, vrlo rijetko koristi.

Ulaz: $\mathcal{G}_1, \dots, \mathcal{G}_m$.

Inicijaliziraj skup trenutnih klastera $C := \{\{\mathcal{G}_1\}, \dots, \{\mathcal{G}_m\}\}$.

za $i := 1, \dots, m - 1$ **čini**

 Nađi klastera $P, Q \in C, P \neq Q$, takve da je

$$d(P, Q) = \min\{d(P', Q') : P', Q' \in C, P' \neq Q'\}.$$

 Stvori novi klaster $C_i := P \cup Q$.

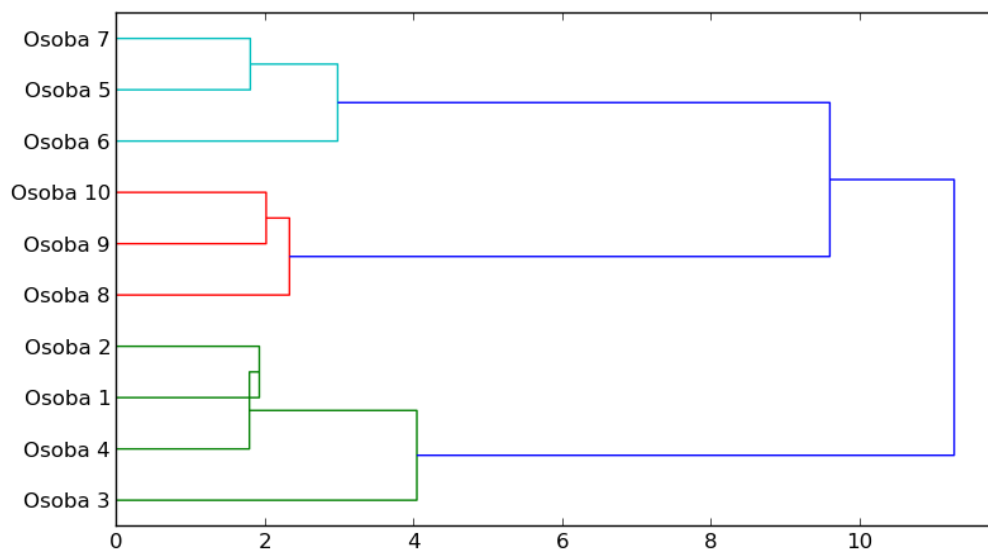
 Spremi $\text{rastav}(C_i) = (P, Q)$.

 Ažuriraj skup trenutnih klastera $C := C \setminus \{P, Q\} \cup \{C_i\}$.

kraj

Izlaz: $\text{rastav}(C_{m-1}), \text{rastav}(C_{m-2}), \dots, \text{rastav}(C_1)$.

Rezultat hijerarhijske klasterizacije možemo zorno prikazati *dendrogramom*, kao na Slici 4.1. Lijevo su pojedinačni objekti, a spajanje dviju linija označava spajanje odgovarajućih objekata ili klastera u viši klaster. x -koordinata spoja odgovara udaljenosti spojenih klastera. Želimo li iz dendrograma dobiti određeni broj međusobno disjunktih klastera, treba ga "prerezati" na određenom nivou (x -koordinati) pa će nezavršene linije odgovarati traženim klasterima.



Slika 4.1: Primjer prikaza hijerarhijske klasterizacije dendrogramom.

4.3 Particijska klasterizacija metodom k -sredina

Za razliku od hijerarhijske klasterizacije, metodu k -sredina (eng. k -means [16]) koristimo za particijsko grupiranje, tj. za dobivanje točno k disjunktnih klastera, za proizvoljno zadani k . Algoritam provodimo s ciljem minimizacije zbroja kvadrata udaljenosti elemenata klastera od središta pripadnih klastera. Preciznije, tražimo particiju $C = \{C_1, \dots, C_k\}$ danog skupa grafova preferencije $S = \{\mathcal{G}_1, \dots, \mathcal{G}_m\}$ koja minimizira funkciju cilja

$$J(C) = \sum_{i=1}^k \sum_{\mathcal{G} \in C_i} d(\mathcal{G}, \mu_i)^2, \quad (4.2)$$

pri čemu je μ_i središte klastera C_i . Središte klastera obično se definira kao centroid, tj. kao aritmetička sredina njegovih elemenata. U našem slučaju, međutim, objekti klasterizacije su grafovi preferencije i njihovo središte moramo drugačije definirati.

U skladu sa razmatranjima iz Poglavlja 2, kao središte klastera C_i uzet ćemo multigraf dobiven agregacijom elemenata klastera. Udaljenost pojedinog elementa klastera od toga središta stoga ćemo prirodno definirati kao udaljenost njihovih potencijala. Dakle,

$$d(\mathcal{G}_i, \mu_j) := \|X^{(i)} - Y^{(j)}\|, \quad (4.3)$$

pri čemu je $X^{(i)}$ potencijal grafa \mathcal{G}_i , a $Y^{(j)}$ potencijal multigrafa μ_j , što je zapravo grupni potencijal klastera C_j .

Algoritam se standardno provodi tako da se početni μ_j nasumično odaberu iz skupa objekata S . U svakoj iteraciji algoritma postoje dvije faze. U prvoj fazi, svaki element iz S pridružuje se njemu najbližem središtu μ_j . U drugoj fazi, iz novodobivenih klastera računamo nova središta μ_j . Iteracije vršimo do konvergencije, tj. dok god se klasteri mijenjaju. Pronađeno rješenje C je lokalni, ali ne nužno i globalni minimum funkcije cilja J . Da bismo povećali vjerojatnost pronalaska globalnoga minimuma i općenito poboljšali rješenje, cijeli algoritam možemo izvrstiti više puta, za različite početne slučajne odabire μ_j , i na kraju ispisati one klastere za koje je vrijednost ciljne funkcije bila najmanja.

Slijedi pseudokod algoritma.

Ulaz: $k, S = \{\mathcal{G}_1, \dots, \mathcal{G}_m\}$.

Izračunaj potencijal $X^{(i)}$ grafa \mathcal{G}_i za $i = 1, \dots, m$.

Na slučajan način odaberi podskup $\{\mu_1, \dots, \mu_k\} \subseteq S$.

$Y^{(j)}$:= potencijal grafa μ_j , za $j = 1, \dots, k$.

ponavljaj

 Za $j = 1, \dots, k$ kreiraj klastere

$C_j := \{\mathcal{G}_i \in S : \|X^{(i)} - Y^{(j)}\| = \min\{\|X^{(i)} - Y^{(j')}\| : j' = 1, \dots, k\}\}$.

 Izračunaj $Y^{(j)}$:= grupni potencijal klastera C_j .

dok klasteri se mijenjaju;

Izlaz: C_1, \dots, C_k .

4.4 Eksperiment

Usporedba parova najlakši je način izražavanja preferencija ([8], [14], [19]) te na najprirodniji način daje grafove prerencije čijom se klasterizacijom bavimo. Međutim, klasterizaciju preferencijskih upitnika s usporedbom u parovima nije lako eksperimentalno evaluirati jer autoru nije poznata nijedna druga metoda klasterske analize takvih objekata koja bi služila za usporedbu.

Srećom, kao što je pokazano u Poglavlju 3, i razni drugi oblici preferencijskih upitnika mogu se pretvoriti u grafove preferencije, a neki od tih preferencijskih upitnika mogu se klasterizirati i standardnim metodama koje nam služe za usporedbu. Jedan su takav slučaj preferencije zemalja u glasanju na Eurosong natjecanju 2005. godine koje smo već koristili u prethodnome poglavlju (Tablica 3.1).

Nije bilo dopušteno glasati za pjesmu iz vlastite zemlje. Ipak, u našoj klsterskoj analizi, radi dobivanja točnijih rezultata, treba uzeti u obzir realnu činjenicu da bi svaka od zemalja koje su imale svoju pjesmu na natjecanju, da je mogla, najviše bodova dala upravo toj pjesmi. Stoga smo pretpostavili da je svaka od tih zemalja vlastitoj pjesmi dala 15 bodova.

Kao i prije, svaka zemlja opisana je grafom preferencije, tj. lancem. Ti su grafovi potom klasterizirani hijerarhijskom analizom, te algoritmom k -sredina za $k = 10$ koji je bio pokrenut 100 puta za različita početna središta μ_j da se pronađe što bolje grupiranje. Algoritmi su implementirani u programskom jeziku Python (za implementacijske detalje vidi Dodatak).

Za usporedbu rezultata, svaku zemlju koja je glasala opisali smo i vektorom – odgovarajućim retkom tablice. Ti su vektori potom klasterizirani standardnom hijerarhijskom analizom te algoritmom k -sredina za $k = 10$, također pokrenutim 100 puta.

U Tablici 4.1 usporedno su prikazani klasteri dobiveni algoritmom k -sredina, u lijevom

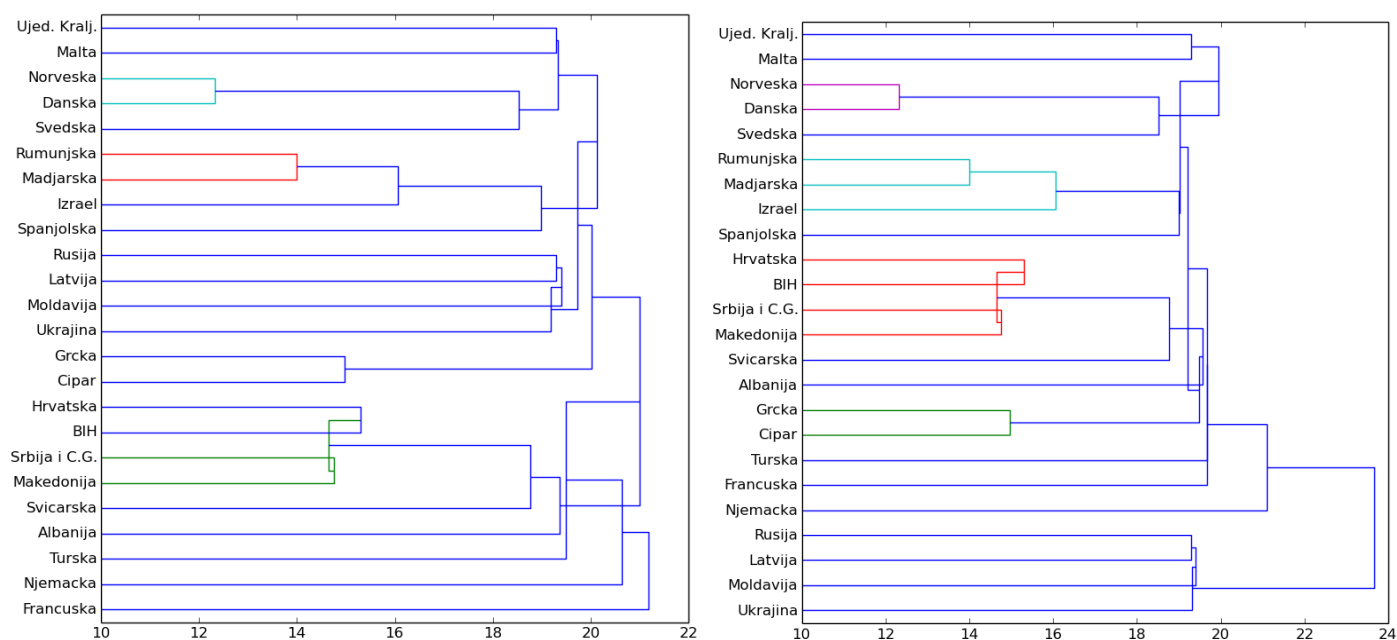
stupcu za predloženi algoritam (klasterizacija grafova preferencije), a u desnom stupcu za vektorski pristup. Stupci su sortirani tako da se uoče podudarnosti. Iz tablice vidimo da predloženi algoritam daje vrlo slične rezultate kao i standardni vektorski algoritam. Također vidimo da dobiveni rezultati odgovaraju očekivanju: susjedne/bliske zemlje imaju slične preferencije pri glasanju.

Na Slici 4.2 dendrogramima su prikazani rezultati naše hijerarhijske klasterizacije (lijevi dio), kao i one vektorskog pristupa (desni dio), i to samo za zemlje čije su pjesme sudjelovale u finalu (24 umjesto 39 zemalja) jer su inače dendrogrami preveliki za vizualno uspoređivanje. Kao i u slučaju k -sredina, vidimo da predloženi algoritam daje slične klasterne kao i standardni vektorski algoritam, te da i ovi rezultati odgovaraju očekivanju: susjedne/bliske zemlje imaju slične preferencije pri glasanju.

Zaključujemo da opisani eksperiment potvrđuje predložene algoritme klasterske analize skupine s pomoću grafova preferencije. Napomenimo još jednom da je eksperiment odabran tako da na njega, radi usporedbe, budu primjenjivi i standardni algoritmi klasterizacije vektora. Oni za većinu preferencijskih upitnika opisanih u Poglavlju 3 nisu primjenjivi, za razliku od ovdje opisane klasterske analize.

Tablica 4.1: Rezultati klusterske analize metodom k -sredina za grafove preferencije (lijevo) i vektore (desno), sortirani tako da se uoče podudarnosti.

koristeći (multi)grafove preferencije	koristeći vektorske profile
Andora Portugal Španjolska	Andora Portugal Španjolska
Danska Island Norveška Švedska Monako	Danska Finska Norveška Švedska
Estonija Latvija Litva Finska Bjelorusija Moldavija Rusija	Island Estonija Latvija Litva Poljska
Ukrajina Poljska	Bjelorusija Moldavija Rusija Ukrajina
Mađarska Izrael Rumunjska Francuska Belgija Nizozemska Turska	Mađarska Izrael Rumunjska
Njemačka	Monako Francuska Belgija Nizozemska Turska Njemačka
Bosna i Hercegovina Hrvatska Slovenija Austrija Švicarska	Bosna i Hercegovina Hrvatska Slovenija Austrija
Bugarska Cipar Grčka	Albanija Bugarska Cipar Grčka
Makedonija Srbija i Crna Gora Albanija	Makedonija Srbija i Crna Gora Švicarska
Irska Malta Ujedinjeno Kraljevstvo	Irska Malta Ujedinjeno Kraljevstvo



Slika 4.2: Usporedni rezultati hijerarhijske klusterske analize za grafove preferencije (lijevo) i vektore (desno). Uočimo da je hijerarhijska struktura klastera u gornjim dijelovima dendrograma (od Ujedinjenog Kraljevstva do Španjolske) potpuno jednaka. Podudaraju se i strukture sljedećih klastera: [Rusija, Latvija, Moldavija, Ukrajina], [Grčka, Cipar], [Hrvatska, Bosna i Hercegovina, Srbija i Crna Gora, Makedonija, Švicarska].

Poglavlje 5

Dodatak

U ovom je dodatku priložen autorov kod Python modula koji sadrži implementaciju klasifikacijskih algoritama za grafove preferencije. Ovaj modul dio je obuhvatnijeg softvera za grafove preferencije koji su napisali L. Čaklović i autor.

```
# Clustering.py

import numpy as np
from random import shuffle
from pyPot.Flow import Flow
from math import sqrt

def flow_distance(flowA, flowB):
    if not flowA.potential:
        flowA.make_potential()
    if not flowB.potential:
        flowB.make_potential()
    ret = 0
    for node in flowA.nodes:
        ret += (flowA.potential[node] - flowB.potential[node]) ** 2
    return sqrt(ret)

def flow_aggregation(flows):
    assert(len(flows) > 0)
    aggr = Flow(nodes=flows[0].nodes)
    for f in flows:
        aggr = aggr.join(f)
    return aggr

def k_means_clusters(flows, k, num_iterations=100, initial_means=None):
    """Returns a list of k lists, each containing flows that belong to a
```

```

particular cluster.
"""
assert(1 <= k and k <= len(flows))
n = len(flows)
best_total_distance = 0

# If initial means are given, it makes no sence to run the algorithm
# multiple times.
if initial_means:
    assert(len(initial_means) == k)
    num_iterations = 1

for iteration in range(num_iterations):
    if initial_means:
        means = initial_means
    else:
        means = flows[:]
        shuffle(means)
        means = means[:k]
    cluster_index = [0 for i in range(n)]
    change = True

    while change:
        # Assignment step.
        change = False
        clusters = [[] for i in range(k)]
        total_distance = 0
        for i in range(n):
            dist_to_closest_mean = flow_distance(flows[i],
                                                means[cluster_index[i]])

            for j in range(k):
                new_distance = flow_distance(flows[i], means[j])
                if new_distance < dist_to_closest_mean:
                    dist_to_closest_mean = new_distance
                    cluster_index[i] = j
                    change = True
            clusters[cluster_index[i]].append(flows[i])
            total_distance += dist_to_closest_mean ** 2

        # Update step.
        for j in range(k):
            if len(clusters[j]) > 0:
                means[j] = flow_aggregation(clusters[j])

# Is this the best clustering found so far?
if iteration == 0 or total_distance < best_total_distance:
    best_total_distance = total_distance

```



```
del representants[j]
del representants[i]

return clusters, linkage_matrix
```

Bibliografija

- [1] C.C. Aggarwal i C.K. Reddy, *Data Clustering: Algorithms and Applications*, Chapman and Hall/CRC, 2013.
- [2] K.D. Bailey, *Sociological classification and cluster analysis*, *Quality and Quantity* **17** (1983), br. 4, 251–268.
- [3] M. Boughanem, Y. Loiseau i H. Prade, *Refining aggregation functions for improving document ranking in information retrieval*, *Scalable Uncertainty Management*, *Lecture Notes in Computer Science* **4772** (2007), 255–267.
- [4] A. Caplin i B. Nalebuff, *Aggregation and social choice: A mean voter theorem*, *Econometrica* **59** (1991), 1–23.
- [5] M. Condorcet, *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*, Imprimerie Royale, Paris, 1785.
- [6] N. Deo, *Graph Theory with Applications to Engineering and Computer Science*, PHI Learning Pvt. Ltd., USA, 2004.
- [7] M.S. Desarkar, S. Sarkar i P. Mitra, *Aggregating preference graphs for collaborative rating prediction*, *Proceedings of the fourth ACM conference on Recommender systems RecSys '10* (2010), 21–28.
- [8] J. Fürnkranz i E. Hüllermeier, *Preference Learning and Ranking by Pairwise Comparison*, *Preference Learning, Part I* (2011), 65–82.
- [9] S. Guha, R. Rastogi i K. Shim, *ROCK: A Robust Clustering Algorithm for Categorical Attributes*, *Information Systems* **25** (2000), br. 5, 345–366.
- [10] J.A. Hartigan, *Clustering Algorithms*, John Wiley & Sons, Inc. New York, NY, USA, 1975.

- [11] N. Mishra, R. Schreiber, I. Stanton i R.E. Tarjan, *Clustering Social Networks*, Algorithms and Models for the Web-Graph, 56–67, 5th International Workshop, WAW 2007, San Diego, CA, USA, December 11-12, 2007. Proceedings.
- [12] T. Pahikkala, E. Tsivtsivadze, A. Airola, J. Järvinen i J. Boberg, *An efficient algorithm for learning to rank from preference graphs*, Machine Learning **71** (2009), br. 1, 129–165.
- [13] M. Pivato, *Aggregation of incomplete ordinal preferences with approximate interpersonal comparisons*, MPRA Paper (2010).
- [14] T.L. Saaty, *Relative measurement and its generalization in decision making: Why pairwise comparisons are central in mathematics for the measurement of intangible factors – the analytic hierarchy/network process*, RACSAM - Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales. Serie A. Matemáticas **102** (2008), br. 2, 251–318.
- [15] T.L. Saaty i L.G. Vargas, *Diagnosis with dependent symptoms: Bayes theorem and Analytic Hierarchy Proces*, Operational Research **46** (1998), 491–502.
- [16] P. Tan, M. Steinbach i V. Kumar, *Introduction to Data Mining. Chapter 8: Cluster Analysis*, Pearson Education, 2005.
- [17] L.L. Thurstone, *A Law of Comparative Judgment*, Psychology Review (1927), br. 34, 273–286.
- [18] M.N. Volkovs i R.S. Zemel, *New learning methods for supervised and unsupervised preference aggregation*, J. Mach. Learn. Res. **15** (2014), 1135–1176.
- [19] H. Wang, Z. Lin. i G. Gediga, *Counting all common subsequences to order alternatives*, Rough Sets and Knowledge Technology, 566–573, 2nd International Conference, RSKT 2007, Toronto, Canada, May 14-16, 2007. Proceedings.
- [20] L. Čaklović, *Conflict Resolution. Risk-As-Feelings Hypothesis*, Labsi Experimental Economics Laboratory (2011), br. 35.
- [21] L. Čaklović, *Teorija vrednovanja s naglaskom na metodu potencijala*, Naklada Slap, 2014.
- [22] L. Čaklović, *Measure of Inconsistency for the Potential Method*, Modeling Decisions for Artificial Intelligence, 102–114, 9th International Conference, MDAI 2012, Girona, Catalonia, Spain, November 21-23, 2012. Proceedings.

- [23] L. Čaklović i A.S. Kurdija, *Universal voting system based on the Potential Method*, preprint (2015).
- [24] M. Šilić, G. Delač i S. Srblić, *Prediction of Atomic Web Services Reliability Based on K-means Clustering*, Proceedings of the 9th Joint Meeting on Foundations of Software Engineering ESEC/FSE 2013 (2013), 70–80.

Sažetak

Usporedba objekata (alternativa) u parovima često se koristi u kontekstu odlučivanja. Metoda potencijala koristi (težinski) graf preferencije kao osnovnu strukturu generiranu takvim usporedbama. Iz toga grafa, rješavanjem sustava jednažbi koji uključuje njegovu Laplaceovu matricu, dobiva se funkcija vrijednosti na skupu alternativa koju nazivamo potencijalom.

U višekriterijskom ili grupnom odlučivanju (npr. izbornim procedurama), svaki kriterij ili sudionik može se predstaviti grafom preferencije. Multigraf dobivem spajanjem tih grafova koristi se za agregaciju preferencija i generira grupni potencijal. Moguće je postaviti proizvoljne težine da bismo podesili utjecaj pojedinog kriterija ili sudionika na grupni potencijal.

Agregaciju grafova preferencije primjenjujemo na izborne procedure. Različiti oblici glasačkih listića generiraju grafove preferencije: tako dobivamo univerzalnu izbornu proceduru koja ne ovisi o obliku glasačkog listića. Još jedna primjena je klasterijska analiza skupine na temelju preferencija njezinih članova, gdje valja provesti hijerarhijsku ili partijsku klasterizaciju više grafova preferencije. U tom kontekstu, agregacijski multigraf koristimo za definiranje središta klastera ili udaljenosti dvaju klastera. Opisanu izbornu proceduru, kao i klasterijsku analizu, za ilustraciju smo primijenili na glasačke preferencije zemalja s Eurosong natjecanja.

Summary

Pairwise comparison of various objects (alternatives) is common in many procedures related to decision making. Potential Method (PM) uses a (weighted) preference graph as the basic structure generated by such comparisons. This graph implies a value function (called potential) on the set of alternatives. The potential is calculated as a solution of a system of equations involving the Laplacian matrix of the preference graph.

In multiple-criteria decision analysis or group decision making (i.e., voting systems), each criterion or decision maker is represented by a preference graph. A multigraph obtained by joining these graphs is used for preference aggregation, generating the group potential. We can set arbitrary weights to adjust the influence of each criterion or decision maker on the group potential.

Aggregation of preference graphs is applied to voting systems. Many different forms of voting ballots can generate preference graphs, which gives us a universal (ballot-independent) voting system. Another application is a cluster analysis of the group based on the members' preferences, where a hierarchical or partitional clustering of their (multiple) preference graphs should be performed. In this context, the multigraph-based aggregation is used to define the center of the cluster or the distance between two clusters. As an illustration, we have applied the described voting system, as well as the cluster analysis, to the voting data from the Eurovision Song Contest.

Životopis

Adrian Satja Kurdija rođen je 21. studenog 1991. godine u Sarajevu, a u Zagrebu živi od ranog djetinjstva, gdje pohađa Osnovnu školu Mladost, a zatim i Petu gimnaziju gdje je proglašen učenikom generacije. Osim čestih nagrada na državnim natjecanjima iz matematike i informatike, osvaja i brojne medalje na međunarodnim olimpijadama iz tih predmeta. Tijekom gimnazije bio je glavni urednik matematičko-informatičkog časopisa *PlayMath*, a 2010. godine jedan od urednika knjige *Čudesna matematika* akademika Vladimira Davidéa. Objavio je ukupno 15 članaka u časopisima *Matka*, *PlayMath* i *Matematičko-fizički list*. Od 2009. do 2012. godine bio je uspješan mentor grupe za matematičke natjecatelje u Petoj gimnaziji, a od 2010. do 2012. godine voditelj radionice za informatičke natjecatelje u Zagrebačkom računalnom savezu. Na Prirodoslovno-matematičkom fakultetu Sveučilišta u Zagrebu 2013. godine postaje sveučilišnim prvostupnikom matematike te upisuje diplomski studij Računarstvo i matematika. Tijekom studija dobiva dvije nagrade Matematičkog odsjeka i nagradu Fakultetskog vijeća PMF-a za izniman uspjeh u studiju. Godine 2011. i 2012. bio je jedan od predstavnika Sveučilišta u Zagrebu na Srednjoeuropskom regionalnom natjecanju u programiranju (ACM - CERC). Godine 2012. u suradnji s docentima Ž. Iličem i G. Šišulom (FER, Zagreb) objavljuje znanstveni rad *Unequal-Slot Based Data Rate Maximization Algorithm for Multiuser OFDMA Systems* u cijenjenom časopisu *IEEE Communications Letters*. Sudjelovao je na tromjesečnim stručnim praksama u softverskim tvrtkama *imo.im* (2012.) i *Google* (2013.). Kao gostujući predavač sudjeluje u pripremama informatičkih natjecatelja u Hrvatskom savezu informatičara te u provedbi vještine Natjecateljsko programiranje na Fakultetu elektrotehnike i računarstva. Bavi se izradom zadataka i test podataka za natjecanja u programiranju: od 2011. do 2013. godine bio je predsjednik Povjerenstva za Hrvatsko otvoreno natjecanje u informatici (HONI – COCI), bio je član Znanstvenog povjerenstva (Scientific Committee) za Central European Olympiad in Informatics – CEOI 2013., a od 2014. član je Državnog povjerenstva za informatička natjecanja. Godine 2015. pod vodstvom doc. dr. sc. Lavoslava Čaklovića piše diplomski rad iz teorije odlučivanja: *Grafovi preferencije i njihova primjena*.