Neke tehnike za traženje pseudo-klika u grafu

Štefan, Rok

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Rok Štefan

NEKE TEHNIKE ZA TRAŽENJE PSEUDO-KLIKA U GRAFU

Diplomski rad

Voditelj rada:
doc. dr. sc. Pavle Goldstein

Zagreb, veljača 2023.

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. _____

Zahvaljujem mentoru doc.dr.sc. Pavlu Goldsteinu na pomoći pri pisanju rada, na razumijevanju i strpljenju u suradnji, na uloženom vremenu i trudu te posebno na entuzijazmu koji mi je pri svakom sastanku prenio.

Hvala mojoj obitelji i prijateljima koji su mi tijekom studija pružili podršku i pomoć kada mi je to bilo potrebno.

Sadržaj

Uvod

Problem traženja klike je, u današnje vrijeme interneta, društvenih mreža i povezanosti, od sve većeg značaja. Algoritam, koji bi bio primjenjiv u znanosti i industriji tj. na velikim skupovima podataka, omogućio bi novi uvid u strukturu podataka te potencijalno pojednostavio neke procese ili pak doveo do novih otkrića.

U ovom radu prikazat ćemo aproksimativni algoritam za traženje pseudo-klika u grafovima. Algoritam ima geometrijski pristup tj. problem zapisan u grafu pretvara u geometrijski problem. To se postiže prelaskom u euklidski prostor u kojem tražimo kuglu koja dobro opisuje vrhove iz klike. Cilj algoritma je da pronađe pseudo-kliku, a u slučajevima kada to ne uspije fokusiramo se na pronalaženje "seed-a" klike.

Rad se sastoji od tri poglavlja. U prvom poglavlju navedeni su pojmovi iz linearne algebre, teorije grafova te neki dodatni pojmovi koji su nužni za daljnje razumijevanje rada. Drugo poglavlje bavi se opisom problema, navodi neke poznate algoritme, objašnjava ideju geometrijskog pristupa te opisuje naš algoritam. Konačno, u trećem poglavlju dani su rezultati našeg algoritma, njihova analiza te uvjeti u kojima se algoritam pokazuje kao uspješan.

Poglavlje 1

Matematički pojmovi

U ovom poglavlju navest ćemo teoreme, definicije, propozicije i napomene iz linearne algebre i teorije grafova uz neke dodatne pojmove. Pojmovi su preuzeti iz izvora [1], [2], [3], [4], [5], [6] i [7].

1.1 Linearna algebra

Definicija 1.1.1. *Neka je* $\mathbb{F}$ *neki skup na kojem su definirane operacije zbrajanja* $+ : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$ *i množenja* $\cdot : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$ *koje imaju sljedeća svojstva:*

1) $\alpha + (\beta + \gamma) = (\alpha + \beta) + \gamma, \ \forall \alpha, \beta, \gamma \in \mathbb{F}$*;*

2) postoji $0 \in \mathbb{F}$ *sa svojstvom* $\alpha + 0 = 0 + \alpha = \alpha, \ \forall \alpha \in \mathbb{F}$*;*

3) za svaki $\alpha \in \mathbb{F}$*, postoji* $-\alpha \in \mathbb{F}$ *tako da je* $\alpha + (-\alpha) = (-\alpha) + \alpha = 0$*;*

4) $\alpha + \beta = \beta + \alpha, \ \forall \alpha, \beta \in \mathbb{F}$*;*

5) $(\alpha\beta)\gamma = \alpha(\beta\gamma), \ \forall \alpha, \beta, \gamma \in \mathbb{F}$*;*

6) postoji $1 \in \mathbb{F} \setminus \{0\}$ *sa svojstvom* $1 \cdot \alpha = \alpha \cdot 1 = \alpha, \ \forall \alpha \in \mathbb{F}$*;*

7) za svaki $\alpha \in \mathbb{F}, \ \alpha \neq 0, \ postoji \ \alpha^{-1} \in \mathbb{F}$ *tako da je* $\alpha\alpha^{-1} = \alpha^{-1}\alpha = 1$*;*

8) $\alpha\beta = \beta\alpha, \ \forall \alpha, \beta \in \mathbb{F}$*;*

9) $\alpha(\beta + \gamma) = \alpha\beta + \alpha\gamma, \ \forall \alpha, \beta, \gamma \in \mathbb{F}$*.*

Tada kažemo da je uređena trojka $(\mathbb{F}, +, \cdot)$ ***polje****, a elemente polja nazivamo skalarima.*

Napomena 1.1.2. *Skup realnih brojeva $\mathbb{R}$ s uobičajenim operacijama zbrajanja i množenja je polje.*

Definicija 1.1.3. *Neka je V neprazan skup na kojem su zadane binarne operacije zbrajanja $+ : V \times V \to V$ i operacija množenja skalarima iz polja $\mathbb{F}$, $\cdot : \mathbb{F} \times V \to V$. Kažemo da je uređena trojka $(V, +, \cdot)$* ***vektorski prostor nad poljem*** *$\mathbb{F}$ ako vrijedi:*

1) $a + (b + c) = (a + b) + c, \ \forall a, b, c \in V$;

2) postoji $0 \in V$ *sa svojstvom* $a + 0 = 0 + a = a, \ \forall a \in V$;

3) za svaki $a \in V$, *postoji* $-a \in V$ *tako da je* $a + (-a) = (-a) + a = 0$;

4) $a + b = b + a, \ \forall a, b \in V$;

5) $\alpha(\beta a) = (\alpha\beta)a, \ \forall \alpha, \beta \in \mathbb{F}, \forall a \in V$;

6) $(\alpha + \beta)a = \alpha a + \beta a, \ \forall \alpha, \beta \in \mathbb{F}, \forall a \in V$;

7) $\alpha(a + b) = \alpha a + \alpha b, \ \forall \alpha \in \mathbb{F}, \forall a, b \in V$;

8) $1 \cdot a = a \cdot 1, \ \forall a \in V$.

Definicija 1.1.4. *Neka je V vektorski prostor nad poljem $\mathbb{F}$.* ***Skalarni produkt*** *na V je preslikavanje $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{F}$ koje ima sljedeća svojstva:*

1) $\langle x, x \rangle \geq 0, \ \forall x \in V$;

2) $\langle x, x \rangle = 0 \Leftrightarrow x = 0$;

3) $\langle x_1 + x_2, y \rangle = \langle x_1, y \rangle + \langle x_2, y \rangle, \ \forall x_1, x_2, y \in V$;

4) $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle, \ \forall \alpha \in \mathbb{F}, \forall x, y \in V$;

5) $\langle x, y \rangle = \overline{\langle y, x \rangle}, \ \forall x, y \in V$.

Napomena 1.1.5. *U $\mathbb{R}^n$ kanonski skalarni produkt definiran je s*

$$\langle (x_1, \ldots, x_n), (y_1, \ldots, y_n) \rangle = \sum_{i=1}^{n} x_i y_i.$$

Definicija 1.1.6. *Vektorski prostor na kojem je definiran skalarni produkt zove se* ***unitarni prostor****.*

Definicija 1.1.7. *Neka je V unitaran prostor.* ***Norma*** *na V je funkcija $\|\cdot\| : V \to \mathbb{R}$ definirana s*

$$\|x\| = \sqrt{\langle x, x\rangle}.$$

Propozicija 1.1.8. *Norma na unitarnom prostoru V ima sljedeća svojstva:*

1) $\|x\| \geq 0, \ \forall x \in V;$

2) $\|x\| = 0 \Leftrightarrow x = 0;$

3) $\|\alpha x\| = |\alpha|\,\|x\|, \ \forall \alpha \in \mathbb{F}, \forall x \in V;$

4) $\|x + y\| \leq \|x\| + \|y\|, \ \forall x, y \in V.$

Definicija 1.1.9. *Svaka funkcija $\|\cdot\| : V \to \mathbb{R}$ na vektorskom prostoru V sa svojstvima iz propozicije 1.1.8 naziva se norma. Tada $(V, \|\cdot\|)$ zovemo* ***normirani prostor****.*

Definicija 1.1.10. *Norma koja potječe od kanonskog skalarnog produkta na $\mathbb{R}^n$, definiranog u napomeni 1.1.5, dana je formulom*

$$\|(x_1, \ldots, x_n)\| = \sqrt{\sum_{i=1}^{n} |x_i|^2}.$$

Ova norma zove se ***Euklidska norma****.*

Definicija 1.1.11. *Neka je V normiran prostor.* ***Metrika*** *ili* ***udaljenost*** *vektora x i y je funkcija $d : V \times V \to \mathbb{R}$ definirana s*

$$d(x, y) = \|x - y\|.$$

Propozicija 1.1.12. *Metrika na normiranom prostoru ima sljedeća svojstva:*

1) $d(x, y) \geq 0, \ \forall x, y \in V;$

2) $d(x, y) = 0 \Leftrightarrow x = y, \ \forall x, y \in V;$

3) $d(x, y) = d(y, x), \ \forall x, y \in V;$

4) $d(x, y) \leq d(x, z) + d(z, y), \ \forall x, y, z \in V.$

Definicija 1.1.13. *Neka je $X \neq \emptyset$. Svaka funkcija $d : X \times X \to \mathbb{R}$ sa svojstvima iz propozicije 1.1.12 naziva se metrika ili udaljenost. Tada (X, d) zovemo* ***metrički prostor****.*

Definicija 1.1.14. *Neka su* $x = (x_1, \ldots, x_n)$ *i* $y = (y_1, \ldots, y_n)$ *proizvoljni vektori u* $\mathbb{R}^n$. *Metrika na* $\mathbb{R}^n$, *inducirana Euklidskom normom iz definicije 1.1.10, dana je s*

$$d((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

Ova metrika naziva se ***Euklidska metrika****, a prostor* $\mathbb{R}^n$ *zajedno s tom metrikom nazivamo* ***Euklidski prostor****.*

Definicija 1.1.15. *Neka je* (X, d) *metrički prostor. Za proizvoljno* $a \in X$ *i proizvoljan* $r > 0 \in \mathbb{R}$ *skup*

$$K(a, r) = \{x \in X \mid d(a, x) < r\},$$

nazivamo ***otvorena kugla*** *u X, sa centrom a i radijusom r.*

Napomena 1.1.16. *U Euklidskom prostoru* $\mathbb{R}^n$ *otvorena kugla sa centrom* $a \in \mathbb{R}^n$ *i radijusom* $r > 0 \in \mathbb{R}$ *dana je s*

$$K(a, r) = \left\{ x \in \mathbb{R}^n \;\middle|\; \sqrt{\sum_{i=1}^{n} (a_i - x_i)^2} < r \right\}.$$

Definicija 1.1.17. *Za prirodne brojeve m i n, preslikavanje*

$$A : \{1, 2, \ldots, m\} \times \{1, 2, \ldots, n\} \to \mathbb{F}$$

naziva se ***matrica*** *tipa* (m, n) *s koeficijentima iz polja* $\mathbb{F}$.

1.2 Teorija grafova

Definicija 1.2.1. ***Graf*** *G je uređeni par* (V, E), *gdje je V skup* ***vrhova****, a E skup* 2*-podskupova od V, koje zovemo* ***bridovi****.*

Napomena 1.2.2. *Katkada gornju definiciju proširujemo tako da dopustimo* ***petlje*** *(bridove koji spajaju vrh sa samim sobom),* ***višestruke bridove*** *(više bridova između para vrhova) i* ***usmjerene bridove*** *(bridovi koji imaju orijentaciju tako da idu od jednog vrha prema drugome). Usmjereni bridovi se reprezentiraju uređenim parovima, a ne* 2*-podskupovima dok kod višestrukih bridova E postaje multiskup.*

Napomena 1.2.3. *U ovom radu od interesa će nam biti grafovi koji nemaju usmjerenih bridova, petlji niti višestrukih bridova te ćemo u skladu s time tretirati gornju definiciju.*

Definicija 1.2.4. *Kažemo da su vrhovi $u, v \in V$, u grafu $G = (V, E)$,* ***susjedni*** *ako postoji brid $e = \{u, v\} \in E$.*

Definicija 1.2.5. *Kažemo da je graf $G = (V, E)$* ***potpun*** *ukoliko za svaki par vrhova u grafu vrijedi da su susjedni.*

Definicija 1.2.6. ***Podgraf*** *grafa $G = (V, E)$ je graf kojemu su skup vrhova i skup bridova podskupovi od V i E, respektivno.*

Definicija 1.2.7. *Neka je $G = (V, E)$ graf te neka je $|V| = n$. Definiramo* ***matricu susjedstva*** *$A = [a_{i,j}] \in M_{n,n}(\mathbb{R})$, $i, j \in \{1, 2, \ldots, n\}$ sa*

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

Definicija 1.2.8. ***Klika u grafu*** *$G = (V, E)$ je njegov podgraf s bar dva vrha, koji je potpun (tj. postoji brid između svaka dva vrha podgrafa).*

Definicija 1.2.9. ***Maksimalna klika*** *je klika koja nije sadržana u niti jednoj većoj kliki, tj. dodavanjem nekog vrha, ona prestaje biti klika.*

Definicija 1.2.10. ***Najveća maksimalna klika*** *je maksimalna klika koja ima najveći broj vrhova.*

Definicija 1.2.11. ***Slučajan graf*** *je graf $G = (n, p)$ gdje je n broj vrhova, a svaki brid postoji s vjerojatnošću $p \in \langle 0, 1 \rangle$ nezavisno od drugih bridova.*

Teorem 1.2.12. *Neka je $G = (n, p)$ slučajan graf. Tada je* ***očekivana veličina najveće maksimalne klike*** *grafa G jednaka $2 \log_{1/p} n$.*

Definicija 1.2.13. *Neka je $G = (V, E)$ graf i neka su τ i γ t.d. $0 \leq \tau \leq \gamma \leq 1$. Podgraf induciran podskupom vrhova $V' \subseteq V$ je (τ, γ)* ***pseudo-klika*** *ako je*

1) $\forall v \in V'$, $deg_{V'}(v) \geq \tau \cdot (|V'| - 1)$;

2) $|E'| \geq \gamma \cdot \binom{|V'|}{2}$,

gdje je $E' = E \cap (V' \times V')$, a $deg_{V'}(v)$ je broj elemenata s kojima je V' povezan.

1.3 Dodatni pojmovi

Definicija 1.3.1. *Neka je $G = (V, E)$ graf, $|V| = n$ te neka je $A \in M_{n,n}(\mathbb{R})$ pripadajuća matrica susjedstva. Pridružimo mu uređen par (X, d) gdje je $X = V \times V$, a d je zadan kao*

$$d(x, y) = \begin{cases} [A^m(x, y)]^{-1}, & x \neq y \\ 0, & x = y \end{cases},$$

za neki $m \in \mathbb{N}$.

Napomena 1.3.2. *Primijetimo, $d(x, y)$ iz prethodne definicije je zapravo recipročna vrijednost broja puteva duljine m od vrha x do vrha y.*

Napomena 1.3.3. *Uređen par (X, d) iz prethodne definicije nije metrički prostor jer ovako definirana funkcija d ne mora zadovoljavati pravilo trokuta. Stoga, funkciju d nećemo zvati metrikom već **funkcijom različitosti**, a uređen par (X, d) zvat ćemo **skoro metrički prostor**.*

Definicija 1.3.4. *Neka je $G = (V, E)$ graf te neka je $|V| = n$. Definiramo **matricu različitosti** $I = [a_{i,j}] \in M_{n,n}(\mathbb{R})$, $i, j \in \{1, 2, \ldots, n\}$ sa $I_{i,j} = d(i, j)$.*

Definicija 1.3.5. *Ulaganje $f : X \to \mathbb{R}^n$ metričkog prostora (X, d') u euklidski prostor $\mathbb{R}^n$ nazivamo **izometričkim** ukoliko ono čuva udaljenost tj.*

$$\forall a, b \in X \text{ vrijedi } d'(a, b) = d(f(a), f(b)),$$

gdje je d euklidska metrika.

Definicija 1.3.6. *Neka je (X, d) konačni metrički prostor gdje je $X = \{p_0, p_1, \ldots, p_n\}$. Označimo $D_{i,j} = d(p_i, p_j)^2$ te $g_{i,j} = \frac{1}{2}(D_{0,i} + D_{0,j} - D_{i,j})$. **Grammova matrica** metričkog prostora (X, d) je matrica $G = [g_{i,j}] \in M_{n,n}(\mathbb{R})$.*

Teorem 1.3.7. *Konačni metrički prostor (X, d) može se izometrički uložiti u euklidski prostor $\mathbb{R}^n$ ako i samo ako je njegova Grammova matrica pozitivno semidefinitna i ima rang najviše n.*

Teorem 1.3.8. *Neka je $G = (V, E)$ graf, $|V| = n$, $A \in M_{n,n}(\mathbb{R})$ pripadajuća matrica susjedstva te neka je funkcija različitosti vrhova $x, y \in V$ dana kao*

$$d(x, y) = \begin{cases} a, & A(x, y) = 1 \\ b, & A(x, y) = 0 \end{cases},$$

gdje je $0 < a < 1 < b$. Tada postoji $k \in \mathbb{N}$ t.d. je $d_k(x, y) = (d(x, y))^{1/k}$ metrika, a uređen par $(V \times V, d_k)$ metrički prostor.

Teorem 1.3.9. *Neka je $G = (V, E)$ graf, $|V| = n$, $A \in M_{n,n}(\mathbb{R})$ pripadajuća matrica susjedstva te neka je funkcija različitosti vrhova $x, y \in V$ dana kao*

$$d(x, y) = \begin{cases} a, & A(x, y) = 1 \\ b, & A(x, y) = 0 \end{cases},$$

gdje je $0 < a < 1 < b$. Tada postoji $k \in \mathbb{N}$ t.d. je $d_k(x, y) = (d(x, y))^{1/k}$ euklidska metrika, a uređen par $(V \times V, d_k)$ metrički prostor te postoji izometričko ulaganje metričkog prostora $(V \times V, d_k)$ u euklidski prostor $\mathbb{R}^l$ za neki $l \leq n$.

Poglavlje 2

Analiza problema i algoritam

U ovom poglavlju objasnit ćemo problem traženja klike. Opisat ćemo geometrijski pristup egzaktnog traženja klike, navesti najpoznatiji egzaktni algoritam te jedan aproksimativni algoritam. Nakon toga bit će objašnjen naš pristup, ali i proces ulaganja u euklidski prostor. Naposljetku, objasnit ćemo rad našeg algoritma. U ovom poglavlju korišteni su izvori [8], [9] i [10].

2.1 Opis problema i ideja

Koncept pseudo-klike je dobro poznat problem u teoriji grafova. U stvarnom svijetu podatci su rijetko savršeni te je zbog nedovoljne sigurnosti često potrebno dopustiti nekakvu grešku. Iz tih razloga problem traženja klike prirodno je pretvoriti u problem traženja pseudo-klike. Primjena rješenja ovog problema ne nedostaje. Na društvenim mrežama nas zanima koje su to grupe ljudi u kojima se gotovo svi poznaju ili imaju međusobnu interakciju čime čine pseudo-kliku, u bioinformatici nam je od interesa pronaći slične gene koje možemo smatrati pseudo-klikom ako se slično izražavaju u zadanim uvjetima dok u "computer vision-u" pseudo-klikom možemo smatrati piksele koji imaju sličnu boju ili teksturu. U svim ovim slučajevima nam pronalaženje pseudo-klike može pomoći razumjeti strukturu kompleksnih mreža i primijetiti uzorke unutar njih koji na prvi pogled nisu očiti. Problem traženja klike može se riješiti kombinatorno, provjeravajući svaki podskup od skupa vrhova, no vrijeme izvršavanja takvog algoritma je eksponencijalno te u praktičnom smislu taj algoritam postaje beskoristan za velike grafove. Prije nego što objasnimo funkcioniranje našeg algoritma, navest ćemo neke egzaktne algoritme za traženje maksimalnih klika.

Geometrijski algoritam

Opisat ćemo ideju izrade egzaktnoga geometrijskog algoritma. Neka je funkcija različitosti za sve vrhove $x, y \in V$ s pripadajućom matricom susjedstva $A \in M_{n,n}(\mathbb{R})$ definirana kao

$$d(x,y) = \begin{cases} a, & A(x,y) = 1 \\ b, & A(x,y) = 0 \end{cases},$$

za proizvoljne $0 < a < 1 < b$. Tada postoji $k \in \mathbb{N}$ (vidi 1.3.9) takav da redefinirana funkcija različitosti d_k bude metrika, a uređen par $(V \times V, d_k)$ metrički prostor. Ovaj metrički prostor može se izometrički uložiti u euklidski prostor $\mathbb{R}^l$ za neki $l \leq n$. Problem je što je navedeni k jako velik pa nakon ulaganja svi bridovi imaju duljinu skoro jedan te uloženi vrhovi gotovo da postaju n-dimenzionalni simpleks dok vrhovi iz klike čine k-dimenzionalni simpleks gdje je k veličina klike. Sada, problem traženja klike postaje numerički problem jer je pronaći težište kugle koja će dobro opisati vrhove iz klike izrazito složeno.

Bron - Kerbosch algoritam

Bron - Kerbosch je egzaktan algoritam koji pronalazi sve maksimalne klike u zadanom grafu. U svakoj iteraciji algoritam definira skupove R, P i X gdje je R skup vrhova koji su dio trenutne maksimalne klike, P je skup vrhova koji su kandidati za trenutnu maksimalnu kliku, a X je skup vrhova koji nisu dio trenutne maksimalne klike. Na početku algoritma su skupovi R i X prazni dok skup P sadrži sve vrhove. U prvoj iteraciji algoritam uzima vrh v iz P i prebacuje ga u R. U P zadržava samo one vrhove koji su povezani s vrhom v, a ostale prebacuje u skup X. Ovaj proces se ponavlja dok skup P ne postane prazan skup te skup R proglašava maksimalnom klikom. Primijetimo da u svakoj iteraciji u skupu P može postojati više kandidata, a algoritam mora biti izvršen zasebno za svakog kandidata, stoga, možemo reći da je ovaj algoritam zapravo pojednostavljenje gore navedenoga kombinatornog pristupa.

Aproksimativni algoritam

Jedan od poznatijih aproksimativnih algoritama za traženje najveće maksimalne klike je zasigurno onaj iz paketa **NetworkX**, a algoritmu možemo pristupiti pozivanjem funkcije **max_clique**. Ovaj algoritam za dani graf promatra njemu komplementaran graf. U komplementarnom grafu traži najveći nezavisan skup vrhova te u originalnom grafu njega proglašava najvećom maksimalnom klikom.

Pristup

Zbog mnogih prepreka, koje se nameću u rješavanju problema traženja klike, odustat ćemo od nekih pretpostavki što će mnoge prepreke ukloniti. Naravno, ovakav pristup uvodi nove poteškoće, no one će biti iznesene u idućem poglavlju uz prednosti ovakvog pristupa. Problem traženja pseudo-klike riješit ćemo geometrijski aproksimativnim algoritmom. Inicijalni problem koji je zapisan u matrici susjedstva uložit ćemo u euklidski prostor te ćemo tako problem svesti na problem traženja kugle unutar koje želimo da se nalazi što više vrhova iz klike, a što manje onih koji nisu iz klike. Da bismo to postigli, izračunat ćemo matricu različitosti, naš skoro metrički prostor uložiti u euklidski prostor te pronaći kuglu koja dobro opisuje vrhove iz klike.

2.2 Prelazak u euklidski prostor

Nužno je postići da vrhovi koji su dio klike budu bliži jedni drugima od vrhova koji nisu u kliki. Drugim riječima, nužno je da funkcija različitosti (vidi 1.3.1) bude manja za vrhove iz klike od funkcije različitosti za vrhove gdje barem jedan od njih nije iz klike. Primijetimo da za matricu različitosti (vidi 1.3.4) vrlo vjerojatno postižemo ovaj uvjet jer će broj puteva duljine m između vrhova koji su u kliki biti veći nego broj puteva između vrhova gdje barem jedan od njih nije iz klike. Intuitivno, očekivano je da vrhovi koji su iz klike pri odabiru puta duljine m imaju veći broj kombinacija dolazaka do određenog vrha u m koraka jer pri tome mogu koristiti bilo koji od vrhova iz klike.

Primijetimo, u matrici različitosti su na dijagonali dodefinirane nule jer želimo da različitost vrhova i i j bude jednaka 0 kada je $i = j$.

MDS

Multidimensional scaling je stohastički algoritam koji zadani metrički prostor ulaže u euklidski prostor proizvoljne dimenzije. Najčešće se koristi za vizualnu reprezentaciju podataka u 2D ili 3D, no u ovom radu ćemo ga koristiti kako bismo pojednostavili problem i imali valjan prostor za traženje otvorene kugle (euklidski prostor). Algoritmu su potrebni podatci o udaljenosti elemenata, koje mogu biti prethodno izračunate ili ih on može samostalno izračunati, a pretpostavke algoritma su da udaljenosti zadovoljavaju svojstva metrike. Primijetimo da ne možemo uvijek savršeno očuvati udaljenosti, no **MDS** će ih očuvati najbolje što može. To se postiže minimiziranjem stres funkcije S koja je definirana kao:

$$S(x_1, \ldots, x_n) = \sqrt{\sum_{i \neq j = 1, \ldots n} (d_{i,j} - \|x_i - x_j\|)^2},$$

gdje su $x_1, \ldots, x_n$ podatci nakon ulaganja, $d_{i,j}$ udaljenosti između x_i i x_j u originalnom prostoru, a $\|x_i - x_j\|$ udaljenosti u novom prostoru.

Još jednom ćemo napomenuti da ovaj naš skoro metrički prostor nije metrički prostor jer funkcija različitosti nije metrika, no, kao što je već rečeno, dopuštamo određenu nepreciznost s ciljem jednostavnog pronalaska kugle koja dobro opisuje vrhove iz klike.

2.3 Algoritam

Za danu matricu susjedstva dimenzija $n \times n$ konstruiramo matricu različitosti (vidi 1.3.1 i 1.3.4) uzimajući $m = 5$. Pozivanjem funkcije **MDS** iz paketa **sklearn.manifold** naš skoro metrički prostor uložili smo u euklidski prostor $\mathbb{R}^k$ gdje je $k = \lceil \ln n \rceil$, a **MDS**-u su kao udaljenosti dane vrijednosti zapisane u matrici različitosti.

Na ovom euklidskom prostoru preostaje pronaći otvorenu kuglu koja će obuhvatiti što više vrhova iz klike, a što manje vrhova koji nisu u kliki. Za središte kugle uzeli smo težište grafa (n točaka u euklidskom prostoru $\mathbb{R}^k$ koje odgovaraju vrhovima iz grafa), a u prvoj iteraciji algoritma odabrali smo radijus takav da kugla pokupi $\left\lceil 2log_{1/p}\, n \right\rceil$ (što je očekivana veličina klike) vrhova gdje je p proporcija jedinica u gornjem trokutu matrice susjedstva. U svakoj idućoj iteraciji uzimamo po jedan novi vrh sve dok ne postignemo uvjet zaustavljanja.

U svakoj iteraciji promatramo vrhove koji se nalaze u trenutnoj kugli te iz matrice susjedstva uzimamo presjeke onih redaka i stupaca koji odgovaraju vrhovima iz kugle. Time formiramo novu matricu susjedstva (za vrhove iz trenutne kugle) dimenzija $l \times l$.

Uvjet zaustavljanja definiran je preko broja nula u novodefiniranoj matrici susjedstva, a algoritam ponavljamo sve dok je broj nula u gornjem trokutu novodefinirane matrice susjedstva manji od

$$(1 - p) \cdot (l - 1) \cdot log_{1/p}\, n.$$

Ovakav uvjet je odabran jer unutar kugle želimo dopustiti otprilike $log_{1/p}\, n$ (što je polovina očekivane veličine klike) vrhova koji nisu iz klike. To znači da će očekivan broj nula u gornjem trokutu matrice susjedstva dimenzija $l \times l$ biti jednak $(1 - p) \cdot (l - 1) \cdot log_{1/p}\, n$.

Poglavlje 3

Rezultati

3.1 Generiranje matrica

U ovom radu promatrat ćemo grafove veličine 2000. Za prikaz rezultata generirali smo više matrica susjedstva. Kako bismo mogli testirati uspješnost algoritma, u svakoj matrici susjedstva smo kliku veličine k postavili u gornji lijevi ugao dok smo postojanje ostalih bridova simulirali s vjerojatnošću p. Drugim riječima, za prvih k vrhova postavili smo da su svi međusobno povezani dok je svaki drugi par vrhova povezan s vjerojatnošću p. Ovime smo generirali slučajan graf $G = (2000, p)$ uz uvjet da u gornjem lijevom uglu postoji klika veličine k.

3.2 Primjeri

Prikazat ćemo rezultate za vrijednosti $(k, p) \in \{100, 150, 200\} \times \{0.35, 0.50, 0.65\}$. Primijetimo da su za $p = 0.35, 0.50$ i 0.65 očekivane veličine najvećih maksimalnih klika redom $14, 22$ i 35 dok su naše zadane klike znatno veće od očekivanih.

Za svaki uređen par (k, p) generirali smo 7 matrica i na svakoj algoritam ponovili 5 puta. Ponavljanje algoritma je nužno jer je **MDS** stohastičko ulaganje što znači da ne moramo uvijek dobiti iste rezultate. Iako smo 5 puta ponavljali algoritam, zbog jednostavnosti, prikazat ćemo samo 1 rezultat s obzirom na to da u ovih 5 ponavljanja značajnijih razlika u rezultatima nema.

Rezultati su zapisani u obliku uređenih parova (x, y) gdje x predstavlja broj vrhova unutar kugle koji su ujedno i vrhovi iz zadane klike dok y predstavlja broj onih vrhova koji su unutar kugle, ali nisu iz zadane klike. Rezultati su sažeti u iduće 3 tablice gdje svaki redak predstavlja rezultate za navedenu vrijednost p, a svaki stupac rezultate za jednu od 7 različitih matrica koje su nazvane $M_1, M_2, \ldots, M_7$. Primijetimo da su matrice M_i različite za različite vrijednosti p, no zbog kompaktnosti rezultata odabran je ovaj zapis.

Tablica za $k = 200$ i $p = 0.35, 0.50, 0.65$:

p	M_1	M_2	M_3	M_4	M_5	M_6	M_7
0.35	(200, 7)	(200, 8)	(200, 8)	(200, 8)	(200, 8)	(200, 7)	(200, 7)
0.50	(199, 11)	(200, 12)	(199, 12)	(200, 12)	(199, 12)	(200, 11)	(200, 12)
0.65	(187, 19)	(177, 20)	(178, 20)	(181, 20)	(169, 20)	(179, 19)	(170, 19)

Tablica za $k = 150$ i $p = 0.35, 0.50, 0.65$:

p	M_1	M_2	M_3	M_4	M_5	M_6	M_7
0.35	(150, 8)	(149, 7)	(149, 7)	(149, 7)	(150, 7)	(150, 7)	(150, 8)
0.50	(132, 13)	(134, 12)	(110, 12)	(127, 13)	(127, 12)	(136, 12)	(126, 12)
0.65	(105, 21)	(92, 22)	(90, 21)	(103, 21)	(97, 22)	(101, 22)	(97, 22)

Tablica za $k = 100$ i $p = 0.35, 0.50, 0.65$:

p	M_1	M_2	M_3	M_4	M_5	M_6	M_7
0.35	(71, 8)	(73, 8)	(62, 8)	(76, 8)	(78, 8)	(72, 8)	(79, 7)
0.50	(43, 12)	(46, 13)	(38, 15)	(44, 13)	(45, 12)	(46, 15)	(32, 14)
0.65	(29, 26)	(39, 25)	(31, 28)	(36, 27)	(31, 24)	(26, 30)	(39, 24)

3.3 Analiza rezultata

Prije komentiranja rezultata valjalo bi primijetiti da su uspješnost algoritma i k pozitivno korelirani (uz fiksan p) dok su uspješnost algoritma i p negativno korelirani (uz fiksan k). Možemo reći da je ovakav rezultat i očekivan jer su veće klike u grafovima izraženije te ih je lakše pronaći dok veći broj jedinica u matrici susjedstva unosi neku vrstu šuma i otežava pronalazak klike.

Pri tumačenju rezultata nećemo koristiti definiciju pseudo-klike, nego ćemo rezultate tumačiti u praktičnom smislu. Možemo reći da za

$$(k, p) \in \{(200, 0.35), (200, 0.50), (150, 0.35)\}$$

dobivamo skoro cijelu kliku uz nekoliko vrhova koji nisu u kliki te ćemo reći da smo u ovim slučajevima uspješno pronašli pseudo-kliku.

Primijetimo, kada bismo u gore navedenim uređenim parovima fiksirali p i povećali k, također bismo pronašli pseudo-kliku.

Valja napomenuti da za svaki od ovih uređenih parova ukoliko fiksiramo p, a smanjimo k dobivamo (puno) manji skup vrhova iz klike uz nekoliko vrhova koji nisu u kliki te ne možemo reći da smo pronašli pseudo-kliku.

Na prvi pogled se čini da rezultati u ostalim slučajevima nisu zadovoljavajući, no primijetimo da je naša kugla u svim ovim slučajevima ipak uhvatila značajan broj vrhova iz klike te se možemo pitati jesmo li pronašli "seed" klike. "Seed-om" klike nazivat ćemo svaki skup vrhova koji čine kliku, a čiji kardinalitet K zadovoljava sljedeću nejednakost

$$\left(1 - p^K\right)^{2000-K} > 0.9999.$$

Za kardinalitete K koji zadovoljavaju gornju nejednakost možemo reći da smo najmanje 99.99% sigurni da ne postoji vrh koji je s njima povezan kao rezultat slučajnosti nastale generiranjem slučajnog grafa. Kardinaliteti K koji gornju jednadžbu zadovoljavaju za vrijednosti $p = 0.35, 0.5$ i 0.65 su redom $17, 25$ i 39.

Sada, ponovnim gledanjem tablica, možemo reći da smo za gotovo sve kombinacije uređenih parova (k, p) dobili dobre kandidate za "seed" klike, a ponavljanjem istog algoritma na tim vrhovima (uz pripadajuću matricu susjedstva) možemo ukloniti vrhove koji nisu u kliki. Time ćemo izolirati vrhove iz klike te dobiti "seed" klike pomoću kojega stvarnu kliku možemo brzo pronaći iz originalne matrice susjedstva.

U slučajevima gdje nismo dobili kandidate za "seed" klike, konkretno, 5 matrica kada je $(k, p) = (100, 0.65)$, možemo povećati uvjet zaustavljanja te tako pokupiti veći broj vrhova iz klike čime opet dobivamo dobre kandidate za "seed" klike.

Navest ćemo da je vrijeme izvršavanja ovog algoritma otprilike 2 minute uz polinomijalnu složenost te je to njegova glavna prednost nad egzaktnim algoritmima. Ipak, ovakav rezultat dolazi s cijenom toga što algoritam uspješno pronalazi samo one klike koje su znatno veće od očekivanih te su algoritmu zadane povoljne matrice u smislu da sadrže samo jednu veliku kliku.

Kao što smo već naveli, u mnogim situacijama gdje algoritam nije uspješan u pronalasku pseudo-klike, uspješan je u pronalasku "seed-a" klike, a povećanjem uvjeta zaustavljanja u tome postaje još uspješniji. Uvođenje nove klike u graf bi zasigurno unazadilo algoritam, no i u tom slučaju postoji veliki potencijal ovog algoritma uz optimizaciju težišta.

Bibliografija

[1] D. Bakić, *Linearna algebra*, Školska knjiga, Zagreb, 2008.

[2] Š. Ungar, *Metrički prostori*, predavanja, 2016., dostupno na `https://www.mathos.unios.hr/metricki/metricki.pdf`.

[3] D. Veljan, *Kombinatorna i diskretna matematika*, Algoritam, Zagreb, 2001.

[4] N. Alon, J. H. Spencer, *The Probabilistic Method*, JOHN WILEY & SONS, INC., 2000.

[5] M. Brunato, H. H. Hoos, R. Battiti, *On Effectively Finding Maximal Quasi-Cliques in Graphs*

[6] H. Maehara, *Regular embeddings of a graph*, Pacific Journal of Mathematics, 1983.

[7] H. Maehara, *Euclidean embeddings of finite metric spaces*, Elsevier B.V, 2013.

[8] I. Kapec, *Točnost pretraživanja, clustering i klasifikacija*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet (Matematički odsjek), 2021.

[9] V.Kole, *Computing Maximum Cliques in Parallel*

[10] L. Wilkinson, *Multidimensional Scaling*

Sažetak

Ovaj diplomski rad opisuje značaj i problematiku pronalaženja klika u grafovima, a naveli smo i neke poznate ideje i algoritme za traženje klika. Naš algoritam problem koji je zapisan u grafu postavlja kao geometrijski problem u euklidskom prostoru, pronalazi kuglu koja dobro opisuje vrhove iz klike te time pronalazi pseudo-kliku ili "seed" klike ako u prethodnom ne uspije.

Nakon navođenja matematičkih pojmova nužnih za razumijevanje ovog rada, prezentira se algoritam koji u euklidskom prostoru pronalazi kuglu koja dobro opisuje vrhove iz klike. Algoritam je proveden na slučajnim grafovima veličine 2000 uz različite vjerojatnosti bridova i različite veličine klika, a svi grafovi imaju zadanu kliku znatno veću od očekivane.

Za veće klike pseudo-klika je uspješno pronađena dok je u ostalim slučajevima uspješno pronađen "seed" klike. Rezultati pokazuju veliki potencijal ovakvog pristupa te možemo reći da se algoritam pokazao kao uspješan u pronalasku zadane klike čija veličina je znatno veća od očekivane veličine klike. Prikazani algoritam ima iznimno brzo vremensko izvođenje uz polinomijalnu složenost.

Summary

This thesis describes the importance and issues related to finding cliques in graphs. In addition to that, we mentioned some known ideas and algorithms for finding cliques. Our algorithm transforms the problem written in a graph into a geometrical problem in Euclidean space and finds a sphere that well-describes the vertices that belong to the clique. This way we find the pseudo-clique or seed of the clique if the former was not accomplished.

After stating the mathematical terms necessary for understanding this thesis, we present an algorithm that finds a sphere in Euclidean space that well-describes the vertices belonging to a clique. The algorithm was conducted on random graphs of size 2000 with different edge probabilities and different clique sizes. All of the graphs have a previously set clique that is significantly larger than expected.

For bigger cliques, a pseudo-clique was successfully found while in other cases seed of the clique was successfully found. Results show the huge potential for this kind of approach and we can say that the algorithm has shown success in finding the given clique whose size is significantly larger than the expected clique size. Shown algorithm has exceptionally good time performance with polynomial complexity.

Životopis

Rođen sam u Zagrebu 19. ožujka 1995. godine. Školovanje započinjem u Osnovnoj školi Ivana Cankara u Zagrebu nakon koje u istom gradu upisujem X. gimnaziju "Ivan Supek". Po završetku srednjoškolskog obrazovanja upisujem preddiplomski sveučilišni studij Matematika na Prirodoslovno-matematičkom fakultetu Sveučilišta u Zagrebu. Zvanje sveučilisnog prvostupnika matematike stječem 2019. godine te, tada, na istom fakultetu, upisujem diplomski studij Matematičke statistike.

Slobodno vrijeme volim provoditi s obitelji i prijateljima, a hobiji su mi sport, planinarenje i logičke igre.