GRUPIRANJE (KLASTERIRANJE) – PRIMJENA KRUSKALOVOG ALGORITMA

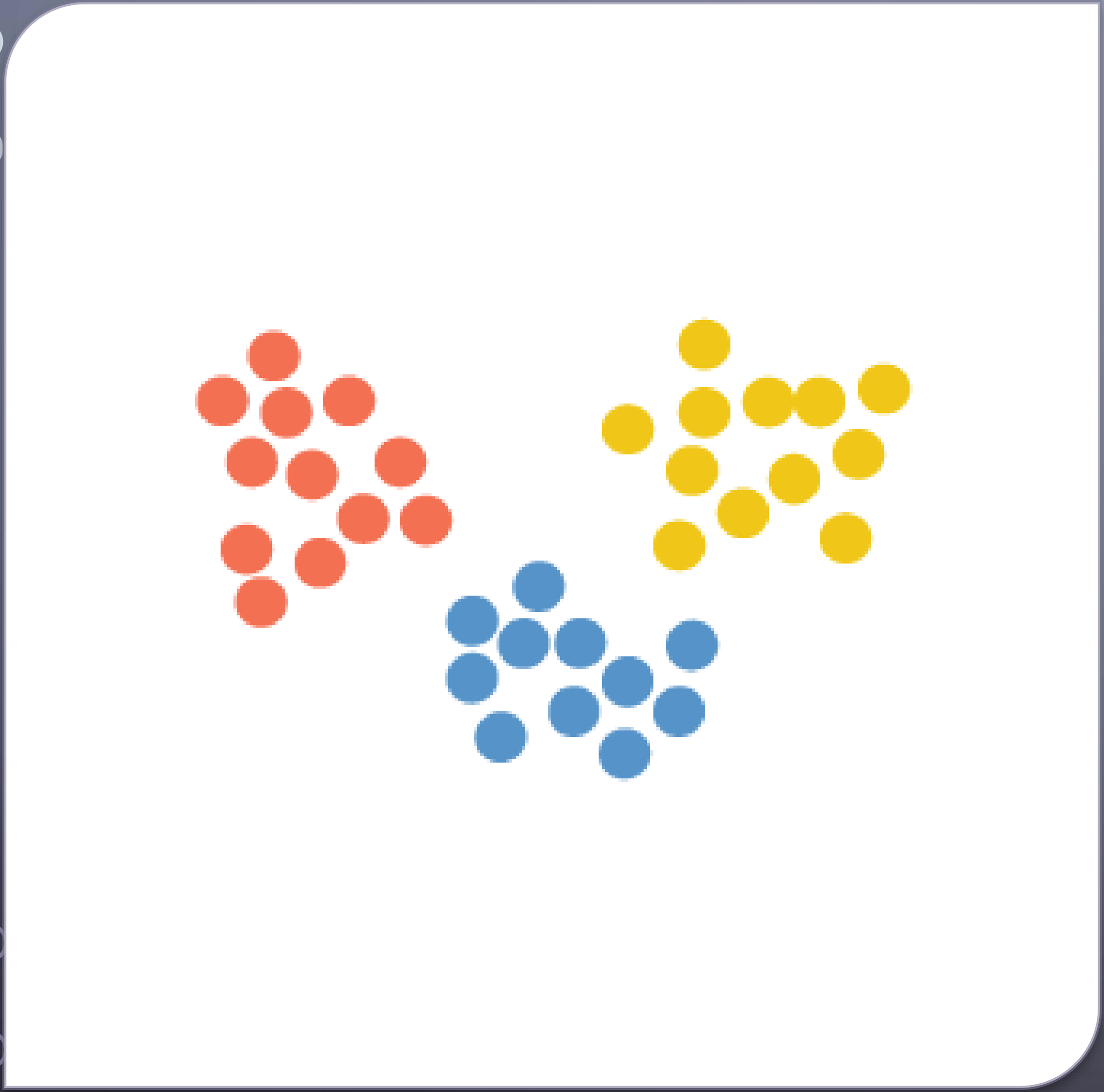
JAKOV KRUNIĆ

SEMINAR IZ KOLEGIJA OBLIKOVANJE I ANALIZA ALGORITAMA

PRIRODOSLOVNO-MATEMATIČKI FAKULTET

ZAGREB, 22.1.2019.

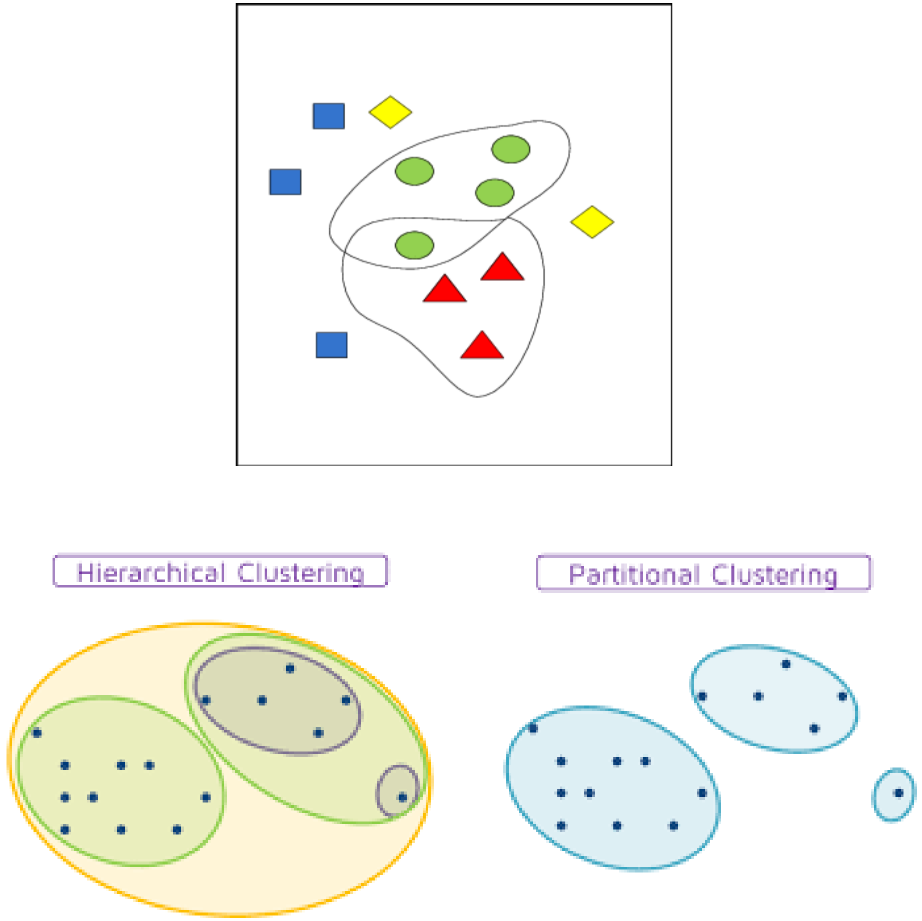
SADRŽAJ



GRUPIRANJE (CLUSTERING)

- Osnovni problem: Neka je S neprazan skup. S treba podijeliti u podskupove (grupe) tako da elementi u svakom podskupu imaju neko zajedničko svojstvo.
- Prirodno je uspostaviti mjeru koja računa koliko je svaki par elemenata sličan.
- Uobičajeni pristup je definiranje funkcije udaljenosti.
- Cilj je da su objekti u istim grupama „bliski”, a objekti u različitim grupama „udaljeni”.

GRUPIRANJE (CLUSTERING)



- Razne opcije grupiranja: hijerarhijsko grupiranje, grupiranje s preklapanjem, strogo razdvajajuće grupiranje s outlierima (ne pripadaju niti jednom klasteru), strogo razdvajajuće grupiranje.
- Proučavamo strogo razdvajajuće grupiranje: svaki objekt pripada točno jednom klasteru.

GRUPIRANJE (CLUSTERING)

- Definicija: Neka je S skup i $k \in \mathbb{N}, k \leq |S|$. **k-grupiranje** od S je particija od S u k nepraznih podskupova.
- Definicija: **Razmak (spacing)** k-grupiranja je najmanja udaljenost između bilo koja dva elementa iz različitih grupa.
- Problem: Za zadani k, treba pronaći k-grupiranje s maksimalnim razmakom.

KRUSKALOV ALGORITAM

- Podsjetnik osnovnih pojmova iz teorije grafova:
  - Graf je uređen par (V, E) pri čemu je V skup vrhova, a E skup bridova (podskup svih dvočlanih podskupova od V).
  - Ciklus je put u grafu koji počinje i završava istim vrhom.
  - Graf je povezan ako između svaka dva vrha postoji put.
  - Stablo je povezan graf koji nema ciklus.
  - Razapinjući podgraf H grafa G je podgraf od G takav da je V(H) = V(G).
  - Promatrat ćemo težinske grafove: svakom bridu je pridružen realan broj (težina).

KRUSKALOV ALGORITAM

- Kako pronaći minimalno razapinjuće stablo?
- Odgovor nam daje Kruskalov algoritam.

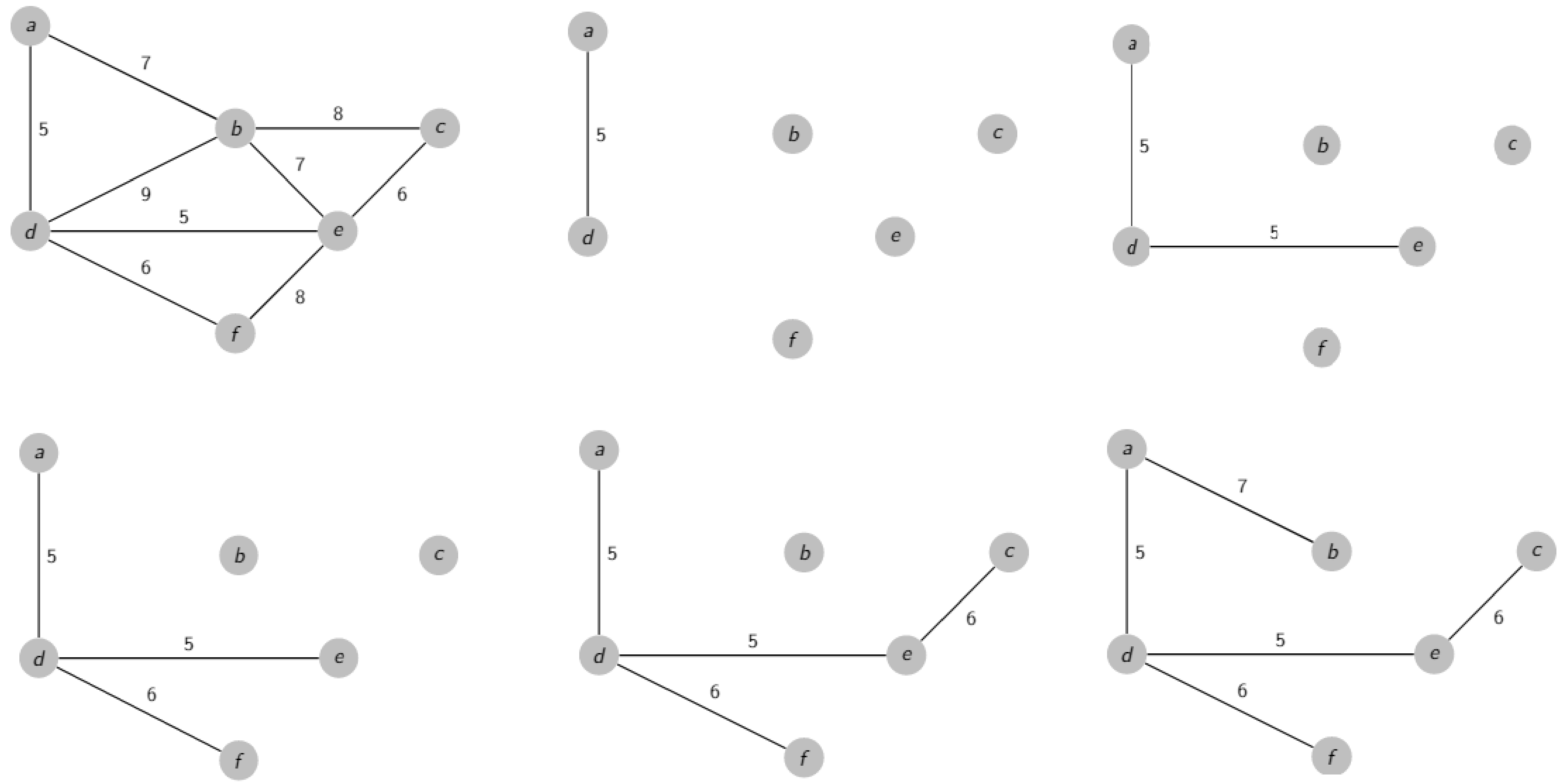
Neka je $G = (V, E)$ povezan težinski graf. Neka je $S = \emptyset$.

Sve dok (V, S) nije povezan ponavljaj:

Odaberi brid $e \in E \setminus S$ minimalne težine takav da $S \cup \{e\}$ nema ciklus

$S = S \cup \{e\}$

- **Teorem:** Kruskalov algoritam nalazi optimalno rješenje.
- **Napomena:** Minimalno razapinjuće stablo ne mora biti jedinstveno.


a
b
c
d
e
f
7
5
8
7
9
5
6
6
8
5
5
5
6
5
5
6
6
7
5
5
6
6

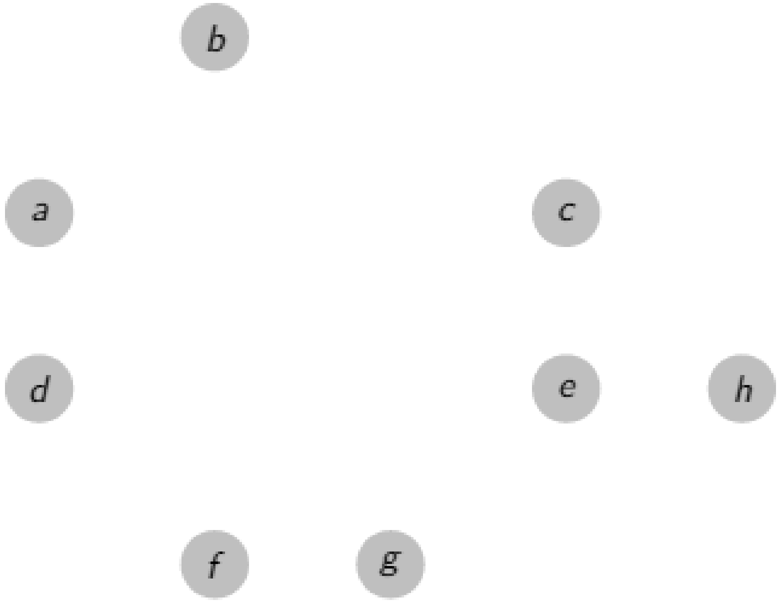
PRIMJENA KRUSKALOVOG ALGORITMA

- Modeliramo problem pronalaska k-grupiranja s maksimalnim razmakom u teoriji grafova.
- Neka je S skup i $k \in \mathbb{N}, k \leq |S|$.
- Izgradit ćemo graf G čiji je skup vrhova S, a komponente povezanosti grafa će biti klasteri.
- Najprije u skup bridova dodamo brid najmanje težine.

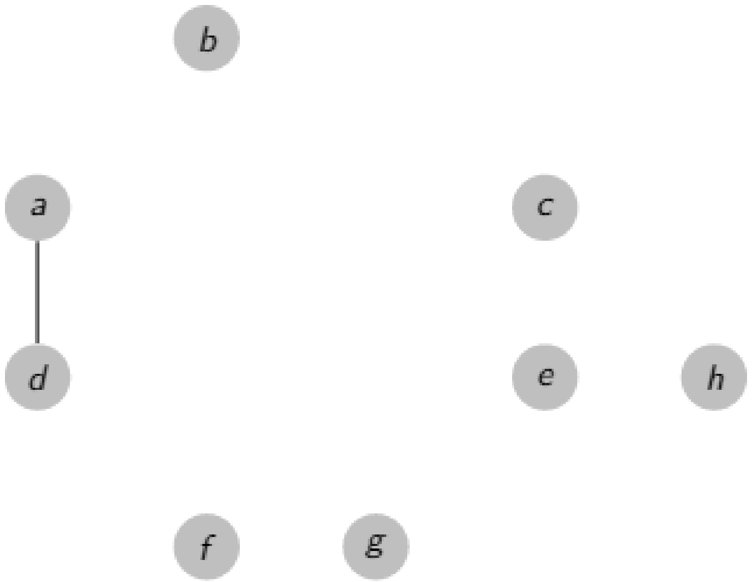
PRIMJENA KRUSKALOVOG ALGORITMA

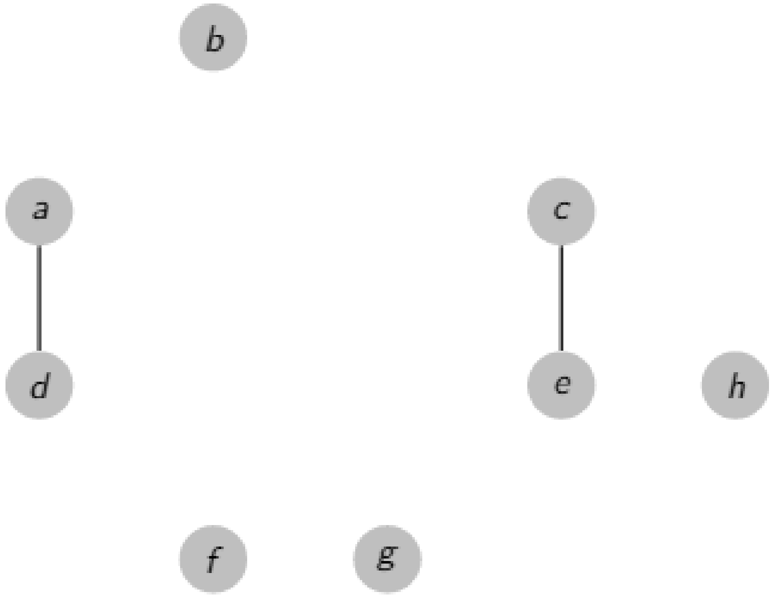
- U svakom koraku biramo brid najmanje težine (od bridova koje još nismo odabrali), uz jedan uvjet.
- Ako je brid (p, q) najmanje težine, a vrhovi p i q se već nalaze u istom klasteru (tj. u istoj komponenti povezanosti), tada taj brid nećemo dodati.
- Na taj način, naš postupak nikad neće stvoriti ciklus, tj. G će biti unija stabala.
- Gotovi smo kad stvorimo točno k komponenti povezanosti.

(a, d)	1	(a, g)	3.5
(c, e)	1	(c, f)	3.5
(e, h)	1	(c, g)	3.7
(c, h)	1	(b, e)	4
(f, g)	1.5	(f, h)	5
(a, b)	1.7	(b, g)	5.5
(b, d)	2	(c, d)	6
(d, f)	2	(d, e)	7
(e, g)	2	(a, e)	7.5
(g, h)	2.5	(b, h)	8
(b, c)	2.5	(b, f)	9
(a, c)	2.7	(d, h)	11
(a, f)	3	(e, f)	11
(d, g)	3	(a, h)	15

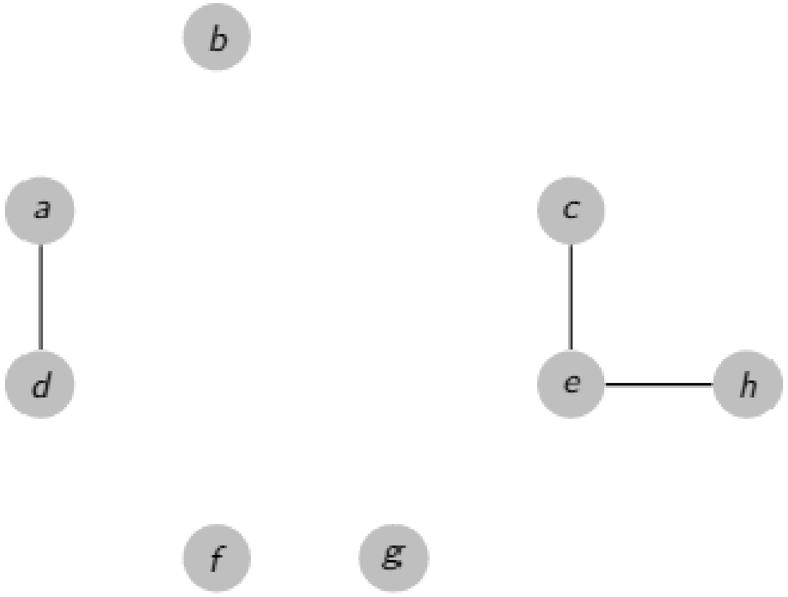


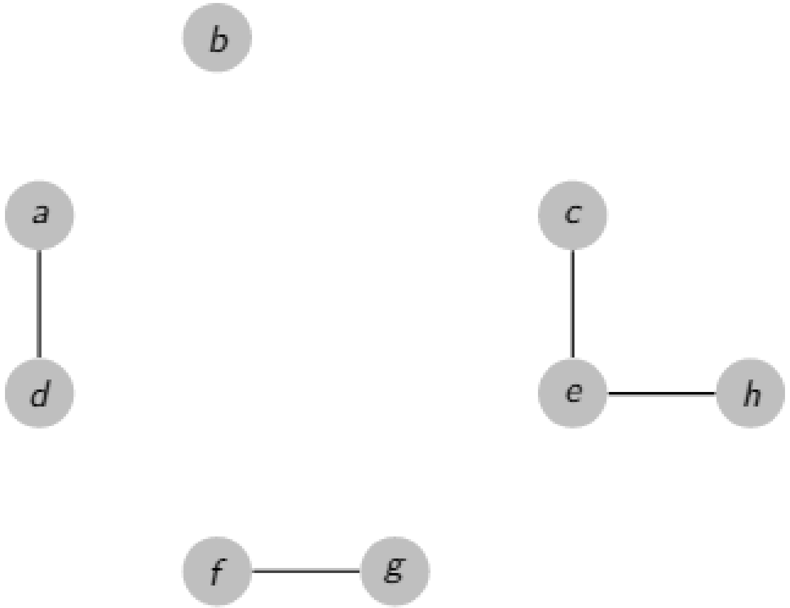
Provedimo postupak za dani primjer i k = 3

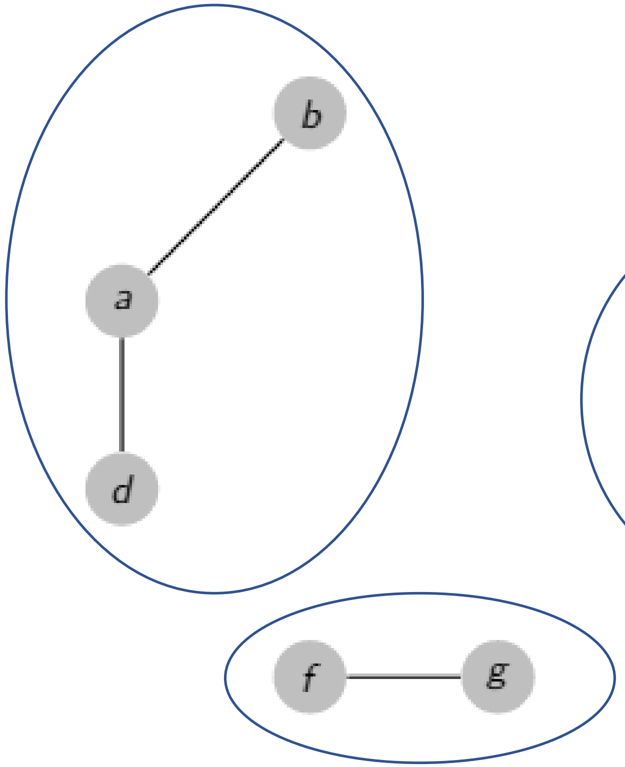




(a, d)	1	(a, g)	3.5
(c, e)	1	(c, f)	3.5
(e, h)	1	(c, g)	3.7
(c, h)	1	(b, e)	4
(f, g)	1.5	(f, h)	5
(a, b)	1.7	(b, g)	5.5
(b, d)	2	(c, d)	6
(d, f)	2	(d, e)	7
(e, g)	2	(a, e)	7.5
(g, h)	2.5	(b, h)	8
(b, c)	2.5	(b, f)	9
(a, c)	2.7	(d, h)	11
(a, f)	3	(e, f)	11
(d, g)	3	(a, h)	15







PRIMJENA KRUSKALOVOG ALGORITMA

- Prethodno opisani postupak je zapravo modificirani Kruskalov algoritam.
- U svakom koraku biramo brid najmanje težine i izbjegavamo cikluse.
- Razlika je u tome što tražimo k-grupiranje, pa zaustavljamo postupak kad dobijemo k komponenti povezanosti.
- Drugim riječima, izvodimo Kruskalov algoritam, ali ga zaustavljamo prije dodavanja posljednjih k-1 bridova.
- Ekvivalentno je da pronađemo minimalno razapinjuće stablo Kruskalovim algoritmom i obrišemo k-1bridova najveće težine.

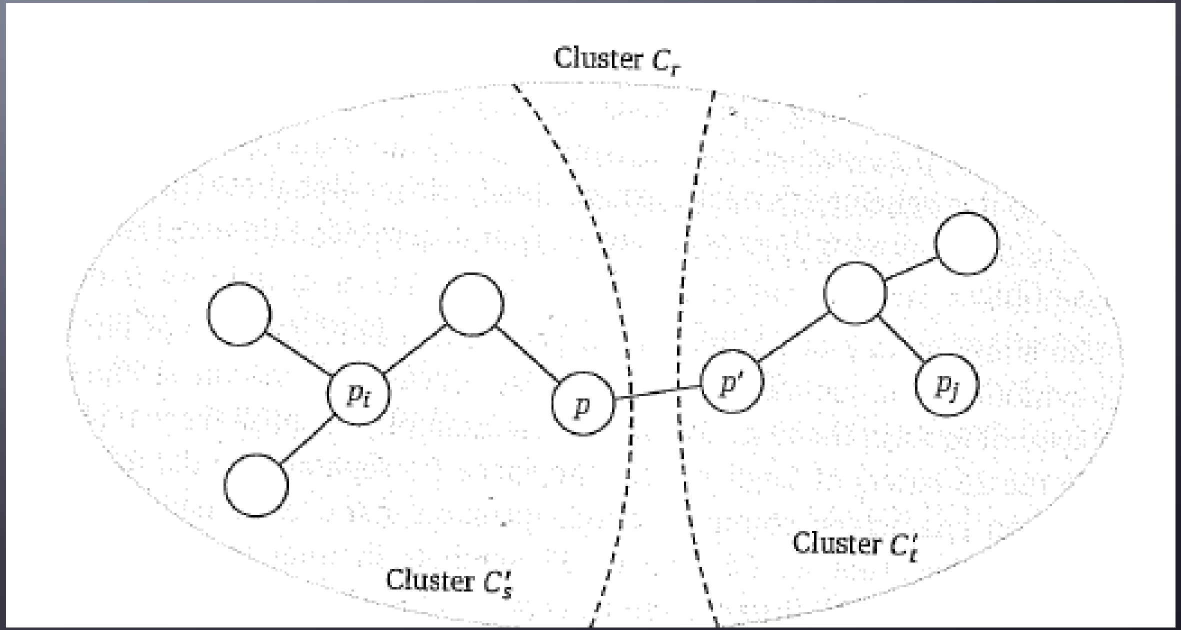
- Jesmo li time dobili k-grupiranje s maksimalnim razmakom?

Teorem

Komponente povezanosti $C_1, C_2, \dots, C_\mathrm{k}$ koje su nastale brisanjem k – 1 bridova najveće težine iz minimalno razapinjućeg stabla čine k-grupiranje s maksimalnim razmakom.

Dokaz: Označimo s C k-grupiranje iz iskaza teorema $C_1, C_2, \dots, C_\mathrm{k}$. Označimo s d^* razmak od C. d^* je zapravo težina (k - 1). najtežeg brida, tj. težina onog brida kojeg bi sljedećeg dodali u naš graf primjenom Kruskalovog algoritma.
Neka je C' proizvoljno k-grupiranje u klastere $C_1{}', C_2{}', \dots, C_\mathrm{k}{}'$; $C' \neq C$.
Trebamo pokazati da je razmak od C' najviše d^*.

Jer je $C' \neq C$, postoji $r \in \{1, \ldots, k\}$ takav da C_{r} nije podskup nijednog ${C_{\mathrm{s}}}'$ iz C'. To znači da postoje vrhovi $p_i, p_j \in C_{\mathrm{r}}$ koji pripadaju različitim klasterima ${C_{\mathrm{s}}}'$ i ${C_{\mathrm{t}}}'$ iz C', tj. $p_i \in {C_{\mathrm{s}}}'$ i $p_j \in {C_{\mathrm{t}}}'$, ${C_{\mathrm{s}}}' \neq {C_{\mathrm{t}}}'$. Jer su $p_i, p_j \in C_{\mathrm{r}}$, postoji put P od p_i do p_j te su svi bridovi iz P nastali primjenom Kruskalovog algoritma. To znači i da je težina svakog brida iz P najviše d^*. Nadalje, znamo da je $p_i \in {C_{\mathrm{s}}}'$ i $p_j \notin {C_{\mathrm{s}}}'$; stoga neka je p' prvi vrh na putu P koji ne pripada ${C_{\mathrm{s}}}'$ i neka je p vrh koji je neposredno prije p' na putu P.



Brid (p, p') je očito nastao primjenom Kruskalovog algoritma pa je njegova težina najviše d^*. p i p' pripadaju različitim klasterima iz k-grupiranja C' i njihova udaljenost je najviše d^*, a razmak k-grupiranja je po definiciji najmanja udaljenost između bilo koja dva elementa iz različitih klastera, pa zaključujemo da je razmak od C' najviše d^*. $\blacksquare$

- Zaključak: Problem k-grupiranja s maksimalnim razmakom rješavamo Kruskalovim algoritmom!

IMPLEMENTACIJA KRUSKALOVOG ALGORITMA

- Sljedeći cilj je efikasna implementacija Kruskalovog algoritma.
- Tražimo strukturu podataka pogodnu za prikaz komponenti povezanosti grafa.
- Cilj je brzo pretraživanje i ažuriranje.
- Struktura podataka bi trebala efikasno spajati dvije komponente u jednu.

STRUKTURA PODATAKA UNION-FIND

- MakeUnionFind(S) uzima skup S i vraća strukturu Union-Find tako da S podijeli na jednočlane podskupove.
- Find(u) vraća „ime” skupa koji sadrži u.
- Union(A,B) spaja skupove A i B u jedan skup.
- Implementacije su moguće pomoću polja i pomoću pointera.

STRUKTURA PODATAKA UNION-FIND

- Složenosti implementacije pomoću polja:
  - MakeUnionFind(S) – O(n)
  - Find(u) – O(1)
  - niz k operacija Union(A,B) – O(k log k)
- Složenosti implementacije pomoću pointera:
  - MakeUnionFind(S) – O(n)
  - Find(u) – O(log n)
  - Union(A,B) – O(1)

IMPLEMENTACIJA KRUSKALOVOG ALGORITMA I VREMENSKA SLOŽENOST

- Promotrimo graf s n vrhova i m bridova.
- Prvi korak je sortiranje bridova uzlazno po težini --> O(m log m).
- Znamo da je $m < n^2$ --> O(m log n).
- Koristimo strukturu Union-Find za prikaz komponenti povezanosti grafa.
- Za brid (u, v) ispitujemo jesu li Find(u) i Find(v) jednaki.

IMPLEMENTACIJA KRUSKALOVOG ALGORITMA I VREMENSKA SLOŽENOST

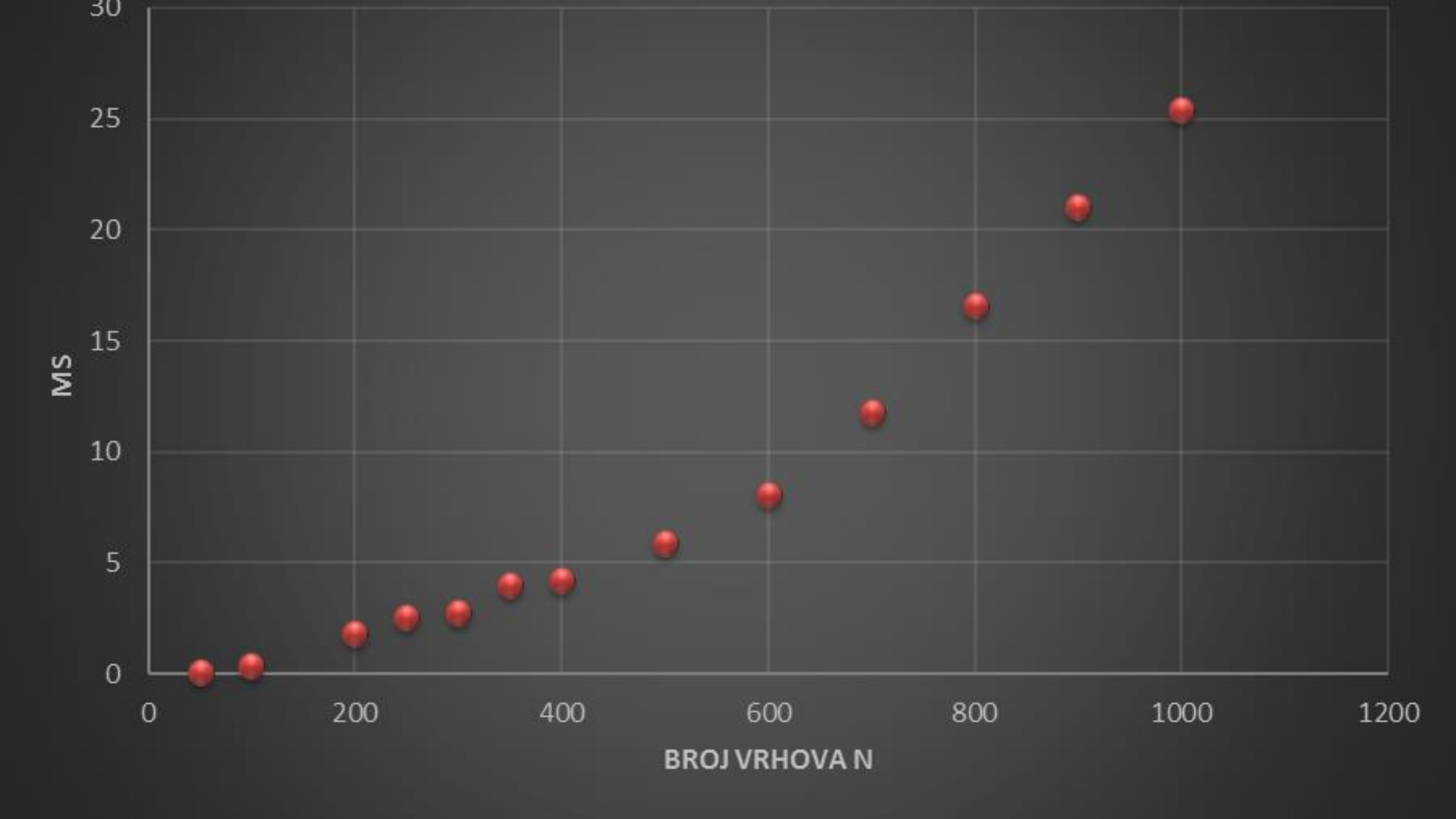
- Union(Find(u), Find(v)) spaja dvije komponente povezanosti.
- Najviše $2m$ operacija Find.
- Točno $n - 1$ operacija Union.
- Ukupna vremenska složenost je O(m log n) i ona ne ovisi o implementaciji Union-Find strukture.

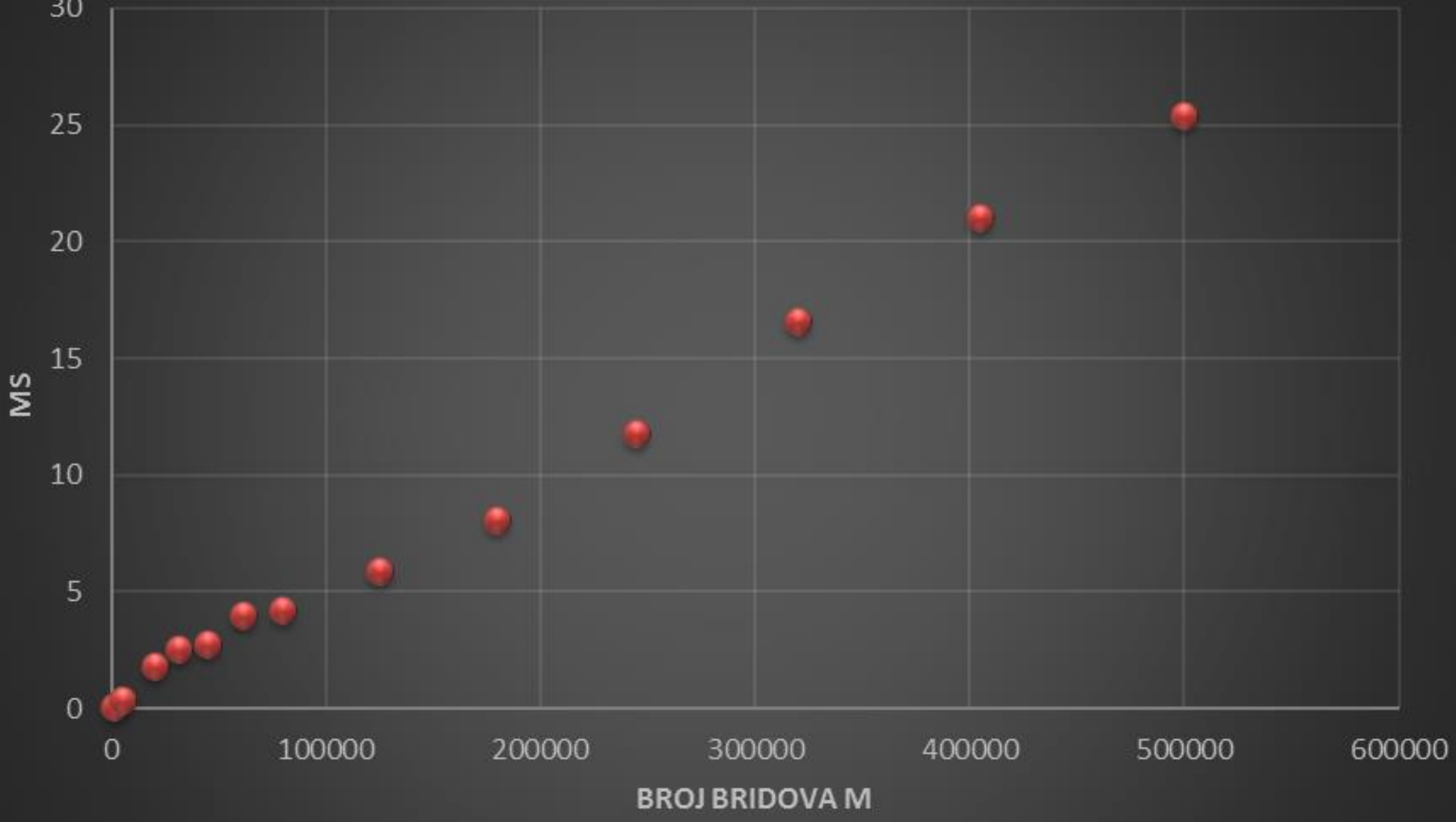
OPIS TESTIRANJA

- Testiramo brzinu Kruskalovog algoritma u ovisnosti o broju vrhova i bridova grafa.
- Ispitujemo vrijeme trajanja Kruskalovog algoritma u milisekundama.
- Uzimamo 100 slučajno generiranih potpunih težinskih grafova s n vrhova.
- Promatramo prosječno vrijeme izvršavanja.
- n biramo između 50 i 1000.

TESTIRANJE

- Za naš početni problem k-grupiranja svi grafovi su potpuni, tj. broj bridova m je jednak $\binom{n}{2}$.
- Očekivana ovisnost o broju vrhova je $n^2 \log n$.
- Uzimamo $k = \frac{n}{2}$.


30
25
20
15
10
5
0
MS
0
200
400
600
800
1000
1200
BROJ VRHOVA N


30
25
20
15
10
5
0
MS
0
100000
200000
300000
400000
500000
600000
BROJ BRIDOVA M

LITERATURA

- P. J. Cameron, *Combinatorics: Topics, Techniques, Algorithms*, Cambridge University Press, Cambridge, 1994.
- A. Karpatne, V. Kumar, M. Steinbach, P.N. Tan, Cluster Analysis: Basic Concepts and Algorithms, dostupno na https://www-users.cs.umn.edu/~kumar001/dmbook/ch8.pdf (8.1.2019.)
- J. Kleinberg, E. Tardos, *Algorithm Design*, Pearson Education, Cornell University, 2006.