

Grafu klasterizācijas problēma un klasteru vizualizēšana

K.Freivalds, P.Ķikusts, Rihards Opmanis, K.Prūsis,
J.Vihrovs

LU konference 2012.gada 11.februāris

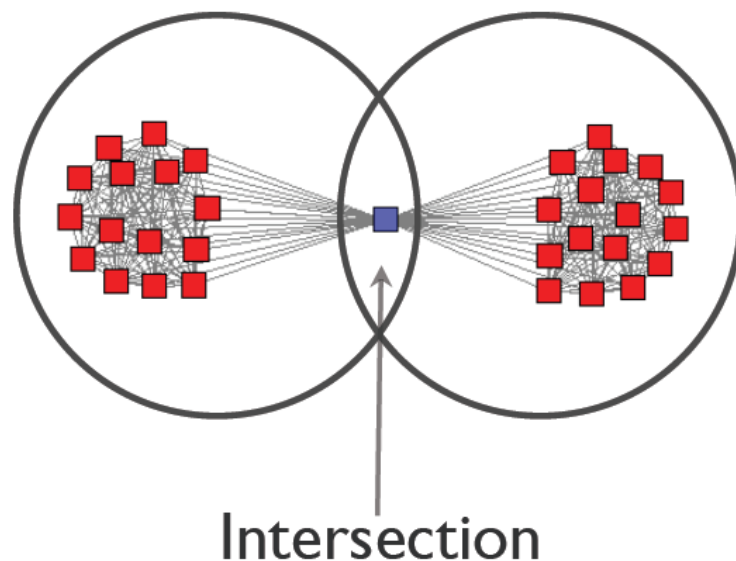
ieguldījums Tavā nākotnē

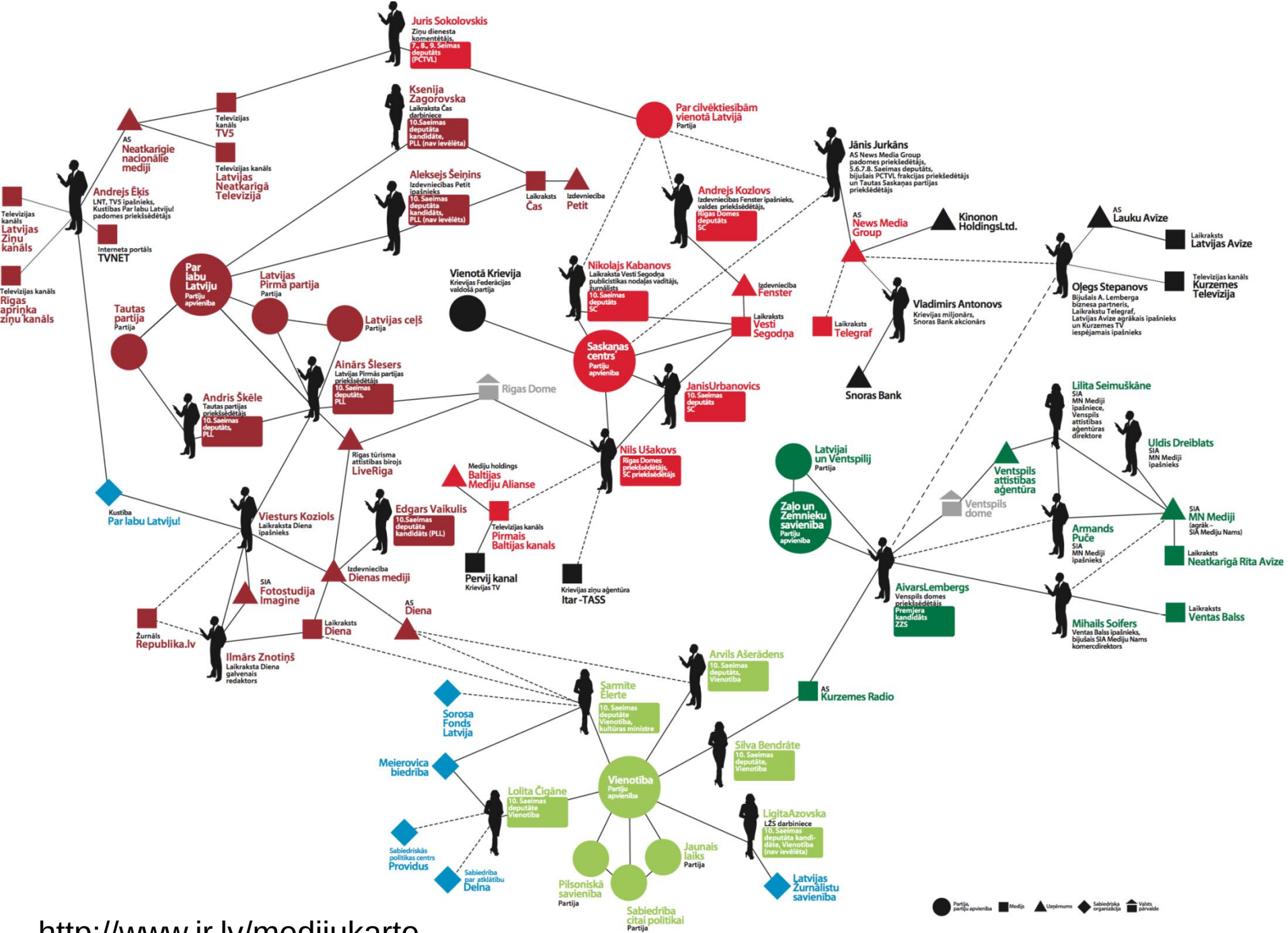
Projekts 2010/0318/2DP/2.1.1.1.0/10/APIA/VIAA/104

Dinamisko tīklu vizualizācijas un analīzes rīku komplekts

Mērķis – grafu sabiezējumu atklāšana un parādīšana

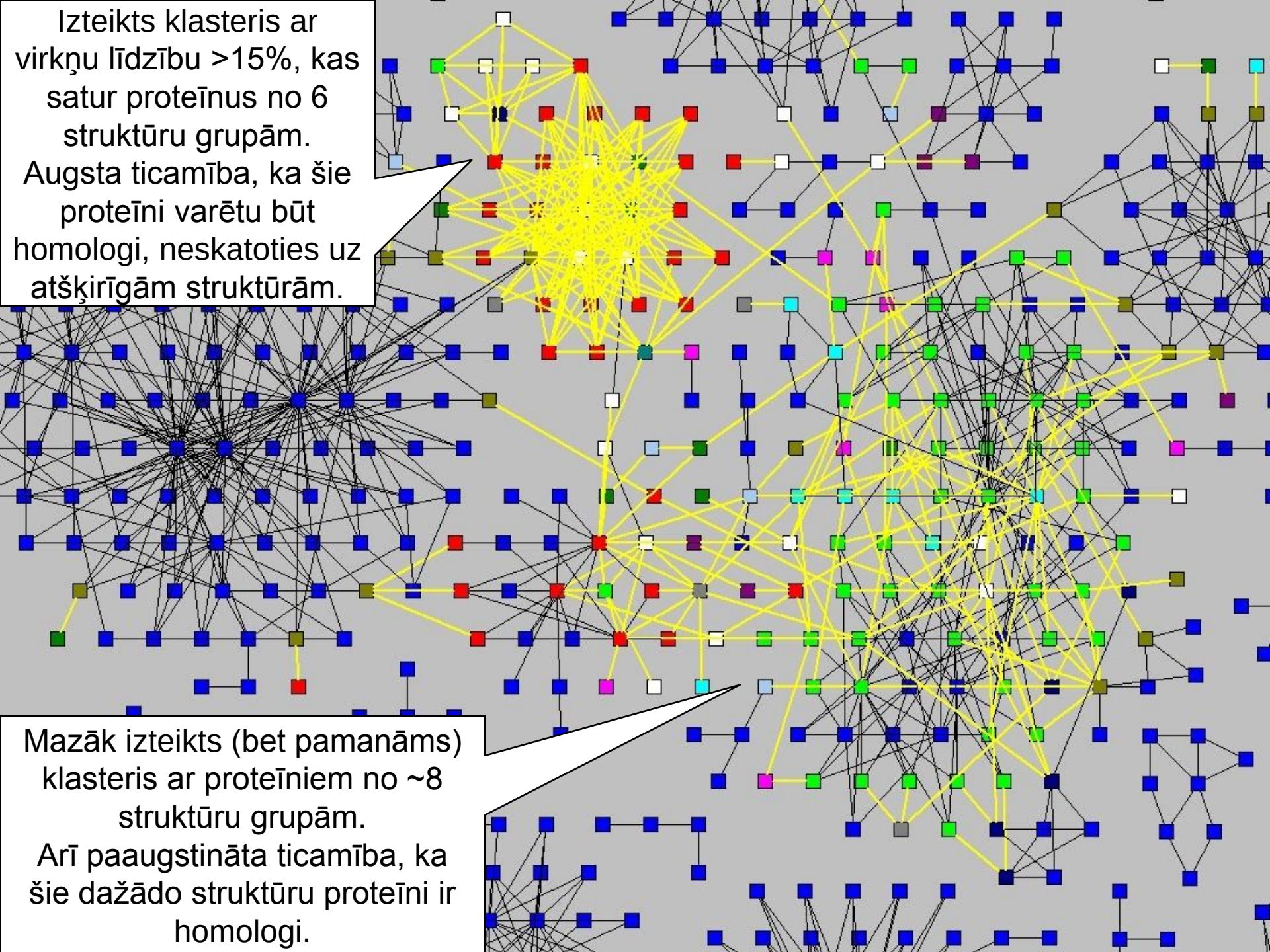
- Sociālo tīklu struktūras atklāšana
 - Klasteri mēdz šķelties
- Noziedzīgu darījumu izsekošana
- Jaunu likumsakarību atklāšana bioinformātikā





Izteikts klasteris ar virkņu līdzību >15%, kas satur proteīnus no 6 struktūru grupām. Augsta ticamība, ka šie proteīni varētu būt homologi, neskatoties uz atšķirīgām struktūrām.

Mazāk izteikts (bet pamanāms) klasteris ar proteīniem no ~8 struktūru grupām. Arī paaugstināta ticamība, ka šie dažādo struktūru proteīni ir homologi.

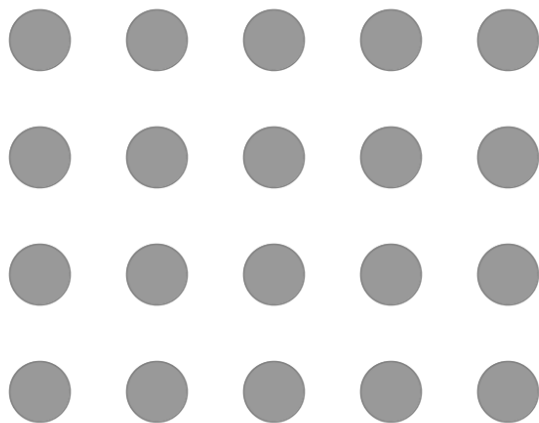


Grafu virsotņu klāsterizācijas problēmas

- Gribam grafu virsotņu klāsterizācijas jautājumu saprast dziļākajā grafu būtībā, pilnīgi atraujoties no iespējamajiem kontekstiem, kuri tika minēti iepriekš.
- Gribētu uzzināt, ar ko tad klāsteri atšķiras no patvaļīga sadalījuma apakšgrafiem.
- Vai tiešām visi virsotņu sadalījumi ir līdzvērtīgi?!

Nelīdzvērtīgu sadalījumu iespējamība

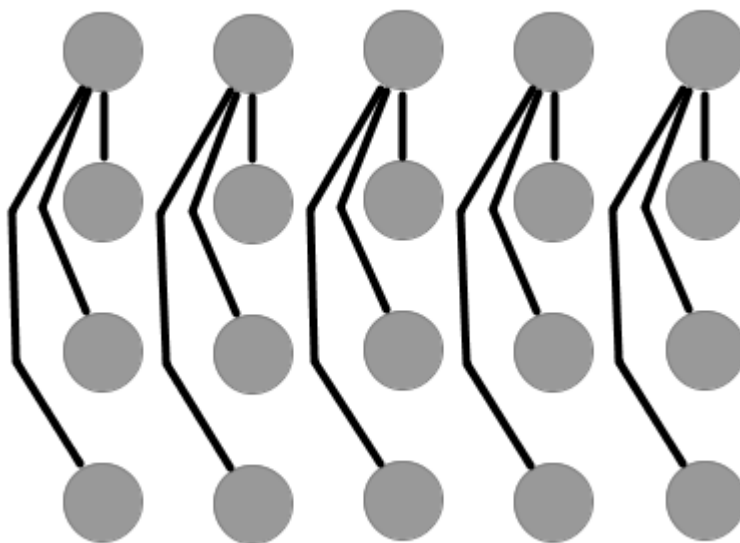
- Mums tiešām nav dabisku kritēriju, lai visvienkāršākā grafa gadījumā, kad vispār nav šķautņu, kādam virsotņu sadalījumam dotu priekšroku.



- Piemēram, virsotņu rindas vai kolonas...

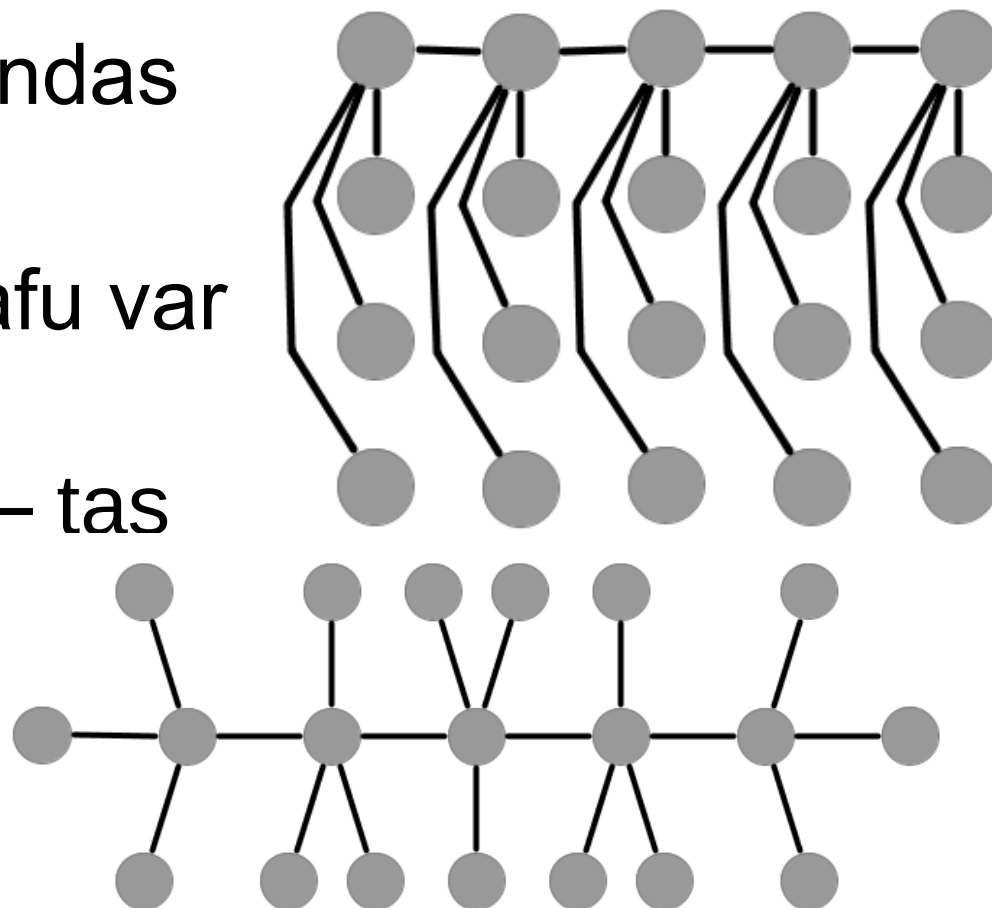
Nelīdzvērtīgu sadalījumu iespējamība

- tomēr, pievienojot kolonās šķautnes, kļūst vismaz grūtāk apgalvot, ka dalījumi rindās vai kolonās ir līdzvērtīgi!

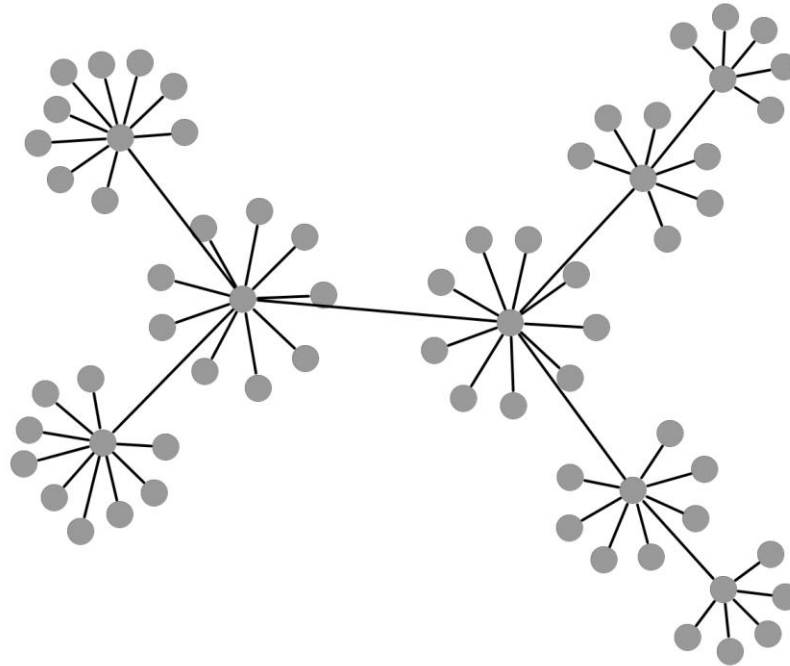


Zvaigžņu koks

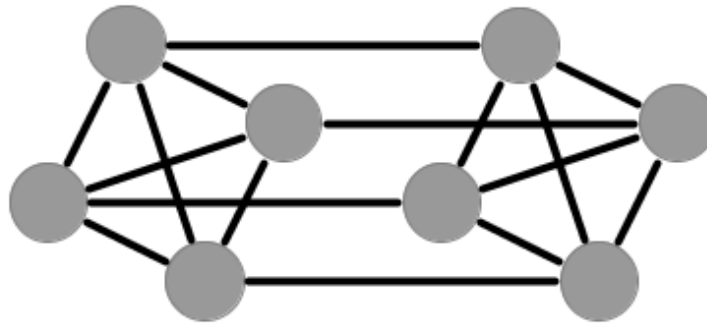
- Uz nupat aplūkotās virsotņu sadalījuma struktūras veidoto grafu var turpināt papildināt, savienojot pirmās rindas virsotnes ar ceļu.
- Tagad arī virsotņu rindas nav līdzvērtīgas
- Savukārt, iegūto grafu var aprakstīt no grafu struktūras viedokļa – tas ir *zvaigžņu koks*.



Grafa struktūra

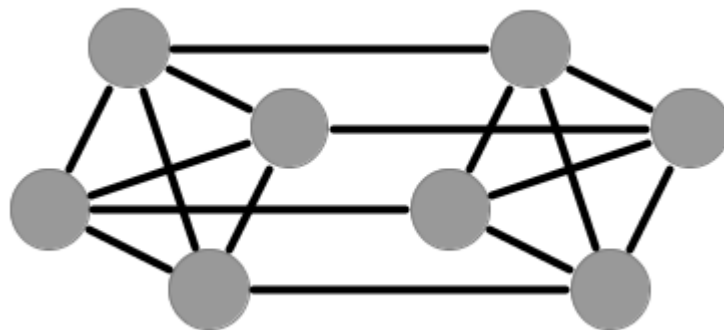
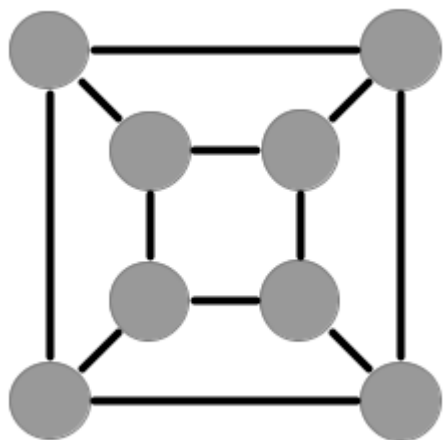


- Zvaigžņu koks piemērs rāda, ka, ja grafs nav pilnīgi homogens, tad tajā ir strukturāli dažādi apakšgrafi, pie tam ar struktūru, kas atšķirīga no grafa vispārīgās struktūras.



Kuba grafa struktūra

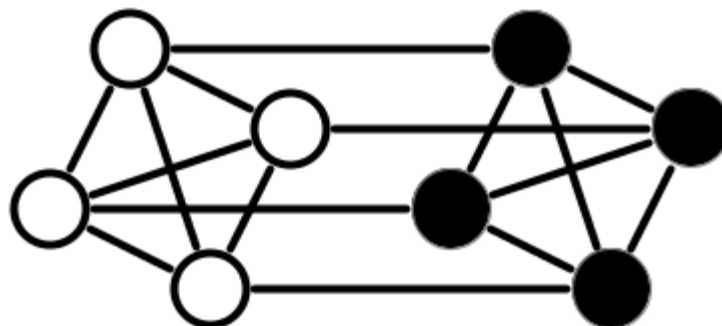
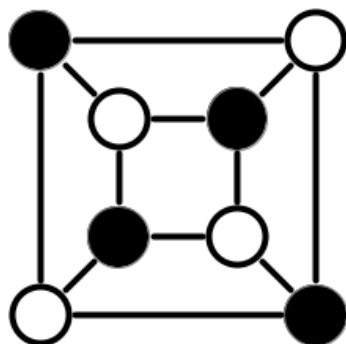
- Apskatīsim vēl vienu, pavisam vienkāršu grafu – kuba. Pievērsīsim uzmanību arī tā papildgrafam.
- Interesanti, ka kuba grafs ir ļoti homogens, atšķirībā no tā papildgrafa.



- Secinājums – īpaši strukturāli izteismīgi ir tukšie un pilnie grafi. Un ja tādi ir, tad tie ir dabiski klasteru kandidāti. Vai kāds domā citādi?

Vai krāsošana ir dabiska klāsterizācija?

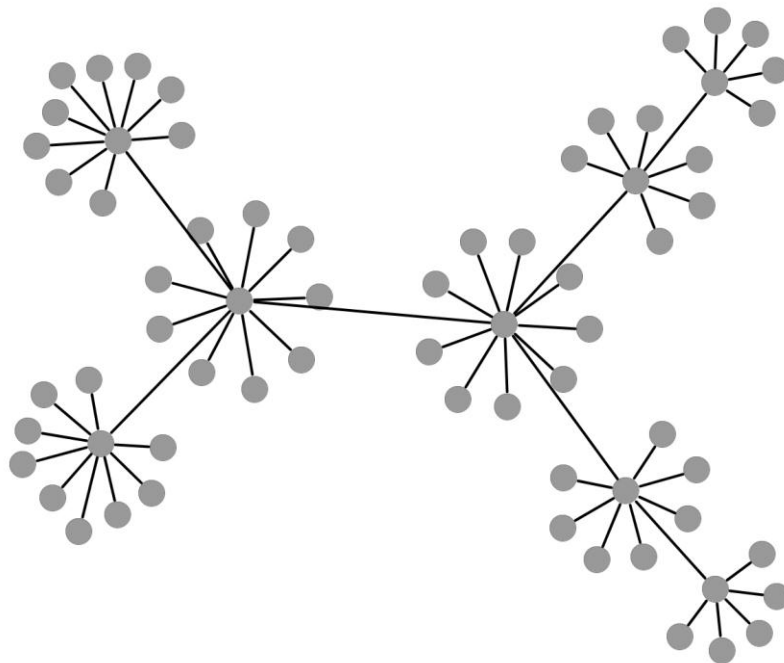
- Mūsu izpētes virzība ir novedusi pie atzinuma, ka grafu virsotņu krāsošana ir īpaši dabiska klāsterizācija.
- Tomēr šāds slēdziens ir krasā pretrunā ar vispāratzīto ideju, ka klāsteri raksturojami kā grafa šķautņu sablīvējumi.



- Tas nozīmē, ka jāturpina iedziļināties pašos jautājuma pamatos – viens virziens te varētu būt atsevišķu grafu klašu izpēte.
- Viena no vienkāršākajām klasēm būtu koki, kuros šķautņu sablīvējumu vispār nav – bet vai tāpēc koku virsotnes nebūtu klāsterējamas jēkā kādā nostādņē, kas kaut cik atbilstu vispāratzītai?!

Vai nāksies mērīt sadalījumu kvalitāti?

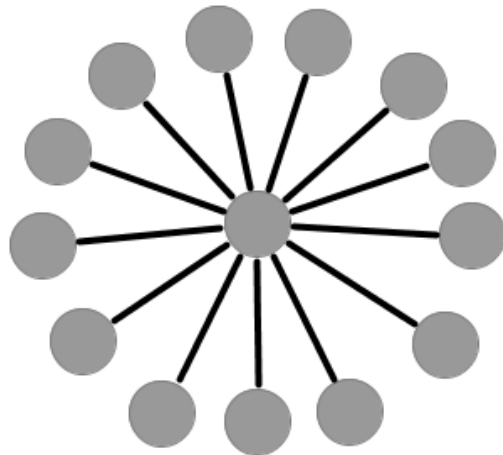
- Zvaigžņu koks ir piemērs tādām kokam, kura struktūras definīcija sniedz arī atbilstošu virsotņu klāsterizāciju – katra atsevišķā zvaigzne ir klāsteris.



- Vai šīs definīcijas diktējums tiešām ir tas, ko vēlamies? Atceroties, ka koki ir divdaļīgi, t.i., to virsotnes it nokrāsojamas divās krāsās, mums ir radusies pretruna ne vien ar vispāratzīto klāsteru blīvuma prasību, bet arī ar mūsu pašu atzīto krāsošanas kā klāsterizācijas dabiskumu.
- Tā mums ir nopietna problēma, kuru joprojām ceram atrisināt, nepievēršoties ārpusgrafa informācijai, t.i., kontekstam.

Vai nāksies mērīt sadalījumu kvalitāti?

- Ekstremāls zvaigžņu koka piemērs ir viena pati zvaigzne – tā uzdod vēl jautājumu, vai vispār tāds grafs būtu jādala klāstēros? Un vai ir jēga interesēties par nedalāmiem grafiem?!



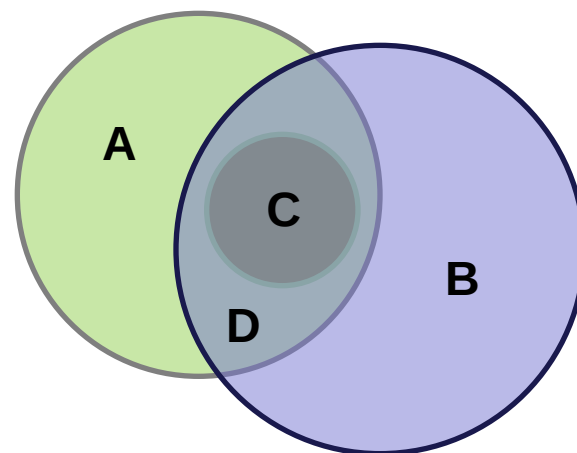
- Uz šādiem jautājumiem mums apmierinošu atbilžu pagaidām nav. Varbūt nāksies samierināties ar domu, ka visi virsotņu sadalījumi ir klasificējami pēc piemērotības tikt atzītiem par klāstērojumiem – literatūrā sen jau ir ievesti dažādi sadalījumu kvalitātes mēri.

Klasteru vizualizācija Eilera diagrammu veidā

- Grafa klasterizācija identificē svarīgas virsotņu kopas, bet nesniedz cilvēkam viegli uztveramu skatu par grafa un tā klasteru struktūru
- Grafa izvietojumam plaknē un iegūtajam klasterējumam tiek veidotas papildus vizuālas struktūras, kas ļauj vieglāk saprast tā saturu

Eilera diagrammas

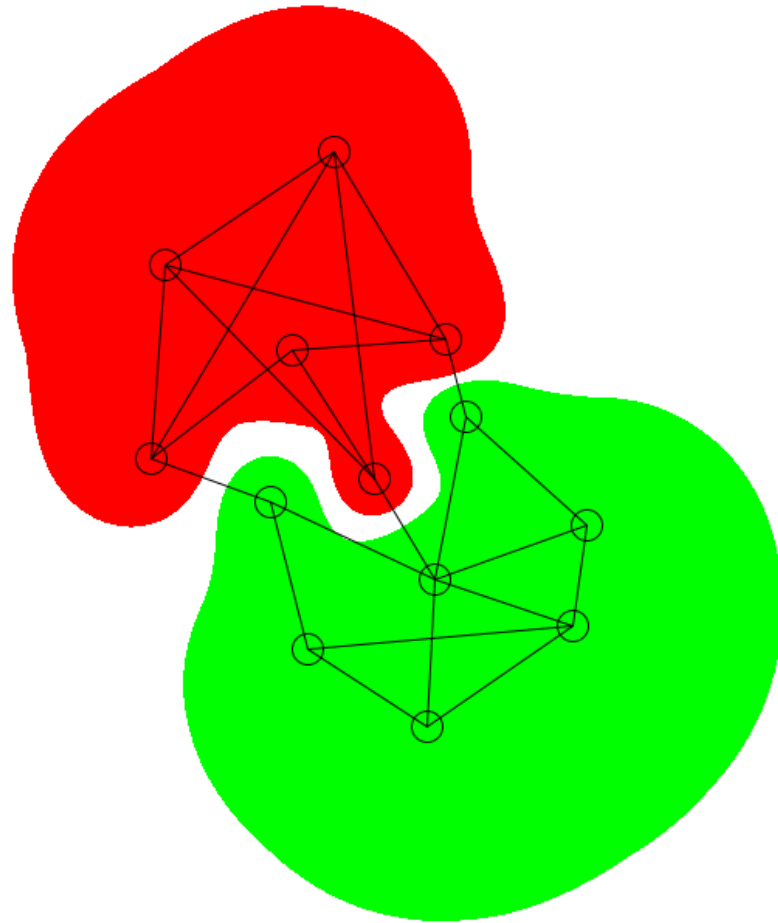
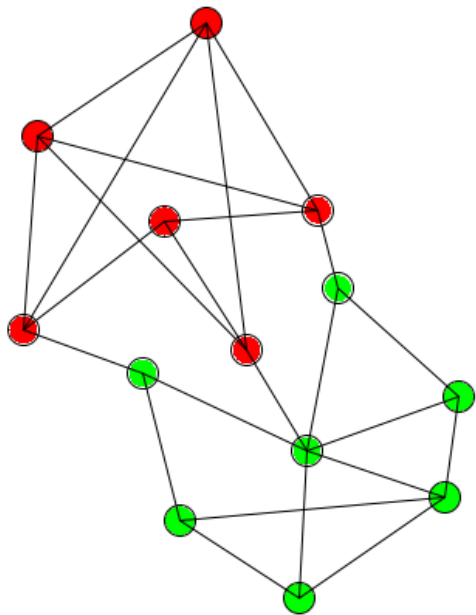
- Tiek lietotas, lai attēlotu kopu savstarpējo struktūru
- Venna diagrammas ir Eilera diagrammu paveids
- Jāizveido Eilera diagrammu no grafa klasteru kopām



Vizualizācija

- Plakne tiek sadalīta apgabalos, katrs tiek iekrāsots sev atbilstošā krāsā
- Katram klasterim atbilst viens apgabals, kurā ietilpst visas šī klastera virsotnes
- Katra virsotne pieder tikai sava klastera apgabalam (ja klasteri nešķēļas)
- Apgabals var būt nesaistīts

Vizualizācija



Vizualizācijas algoritms

Katram plaknes punktam p katram klasterim c tiek rēķināta funkcija f ; punkts piederēs tam klastera apgabalam, kuram šī vērtība būs vislielākā.

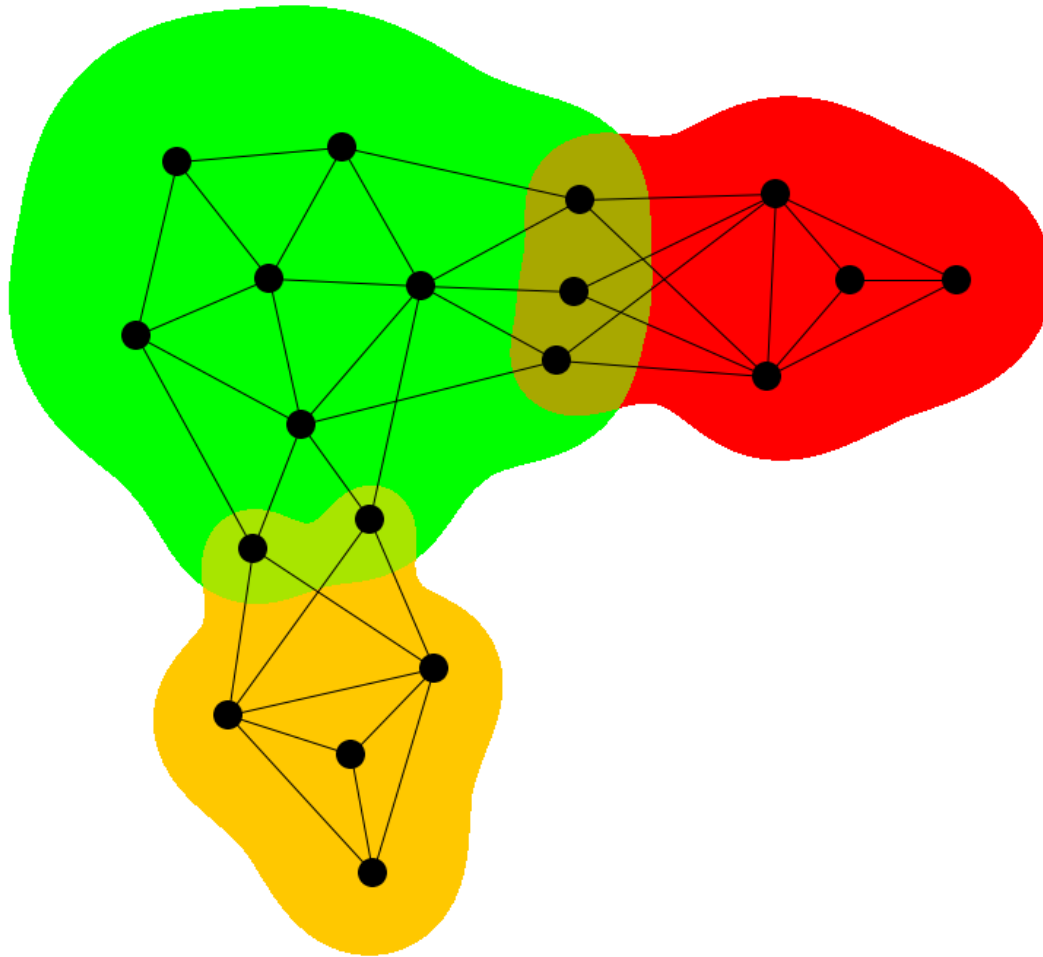
$$f(p, c) = \sum_{i=0}^n \frac{q^i}{a_i} - \sum_{i=0}^m \frac{q^i}{b_i}$$

a_i vērtības ir augoši sakārtoti attālumi no p līdz klastera c virsotnēm; b_i ir augoši sakārtoti attālumi no p līdz pārējo klasteru virsotnēm. q ir koeficients, kas nosaka svaru tālākām virsotnēm no p , $0 \leq q \leq 1$.

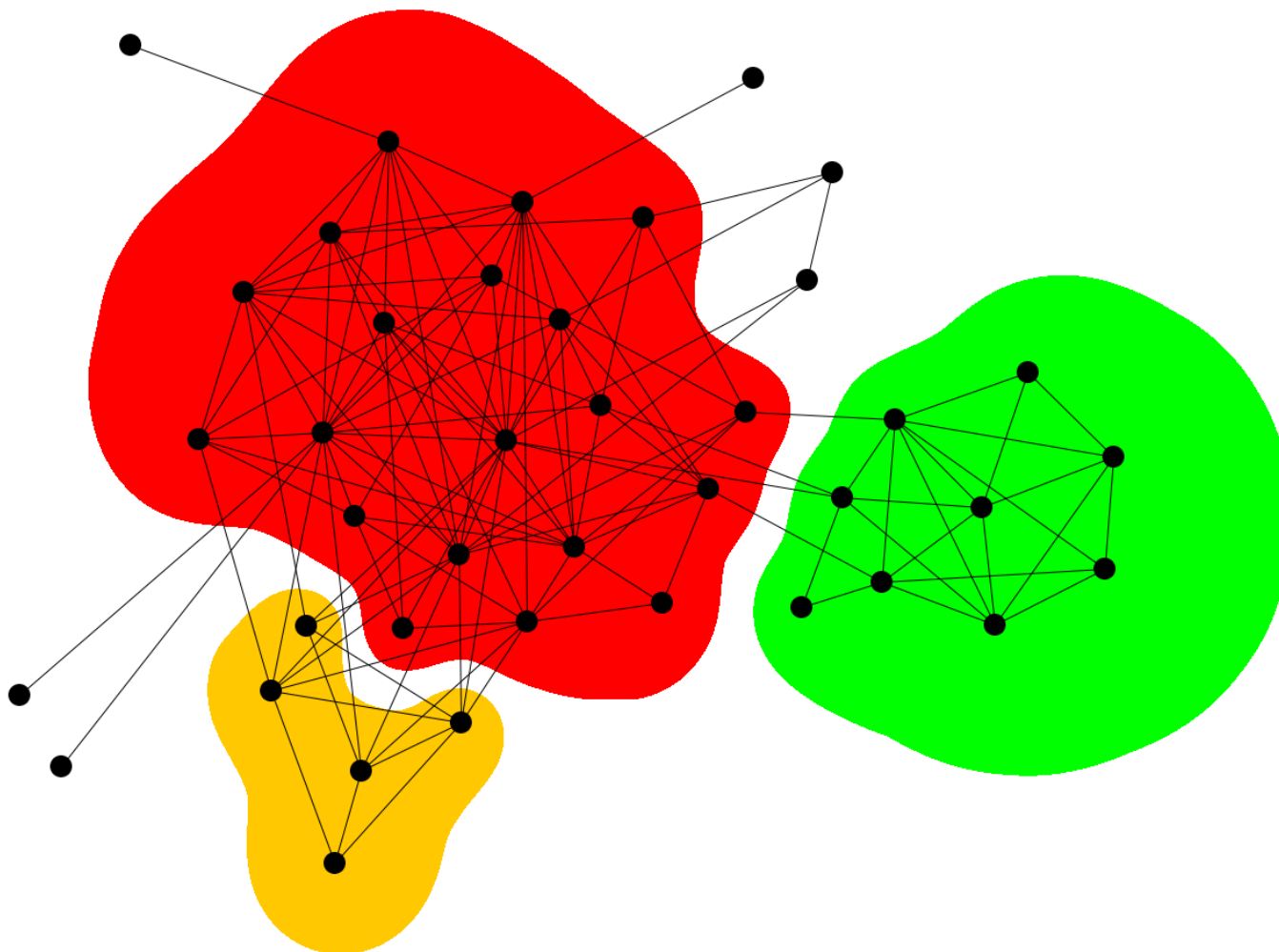
Vizualizācijas rezultāts

- Apgabala robeža ir gluda, ja tam piederošus punktus izvēlas tādus, kuriem f vērtība nav mazāka par kādu konstantu sliekšni un $q \neq 0$
- Ja $q = 0$, tad krāsojums ir precīzi Voronoja diagramma
- Vizualizācija tiek ģenerēta arī grafiem ar šķelošamies klasteriem

Šķelošos klasteru vizualizācija



Reāla sociālā tīkla vizualizācija



Paldies par uzmanību!