

Analýza grafletovej štruktúry reálnych sietí

Diplomová práca

2022

FMFI UK

**autor: Bc. Michal Puškel
školiteľka: doc. RNDr. Mária Markošová, PhD.**

Ciele

- Naprogramovať kombinatorické riešenie počítania orbít grafletov (ORCA) a vytvoriť software na počítanie GDD.
- Vytvoriť software na počítanie a porovnávanie súhlasu GDD sietí.
- Porovnať triviálne rozlíšiteľné umelo vytvorené siete za účelom validácie súhlasu GDD ako vhodnej miery na porovnávanie štruktúry grafov.
- Zistiť, či sa atribút funkčnej siete mozgu “mať Alzheimerovu chorobu” prejavuje v štruktúre jej grafu.

Ciele

Video ukážka aplikácie

- <https://youtu.be/VezPTKmdfNE>

Výsledky

michalpuskel/diplomka: web pa x +

github.com/michalpuskel/diplomka

Aplikácie Bookmarks Pluto Dovolenky 2018/2... MAC

Search or jump to... Pull requests Issues Marketplace Explore

michalpuskel / diplomka Public

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

master 1 branch 0 tags

Go to file Add file Code

About

web page for Master's thesis mAIN FMFI UK

Readme 0 stars 1 watching 0 forks

Releases

No releases published Create a new release

Packages

No packages published Publish your first package

README.md

Graphlet Structure Analysis of the Real Networks

Master's thesis

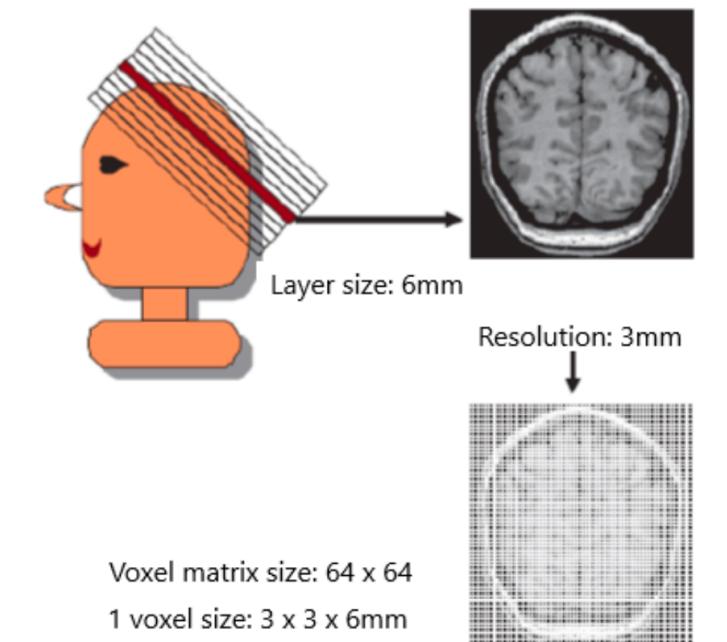
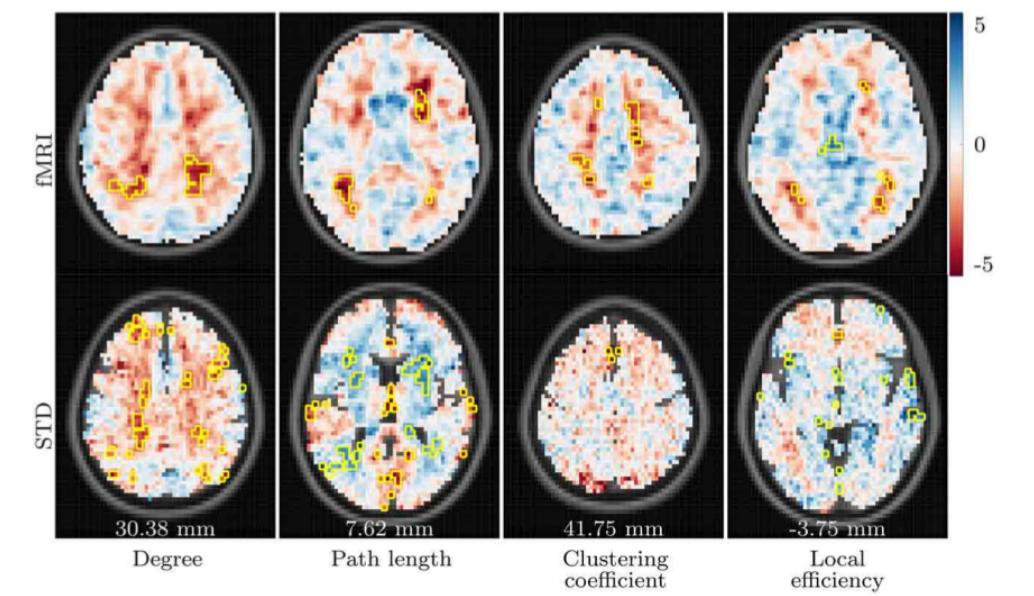
Bc. Michal Puškel
supervisor: doc. RNDr. Mária Markošová, PhD.

2022
FMFI UK

Task description

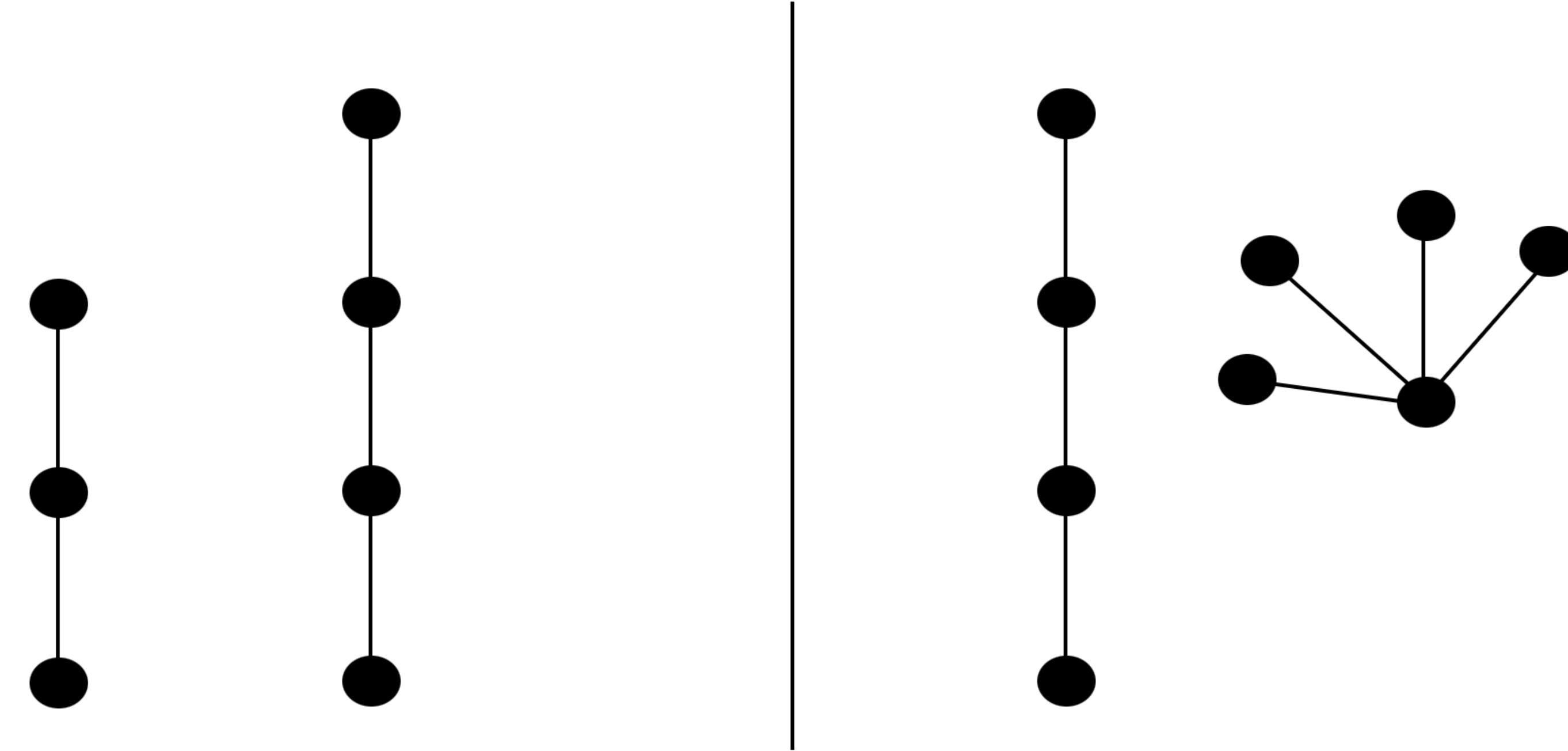
Vstupné dáta

- jednoduché, neorientované, súvislé grafy (zoznam hrán)
- funkčná siet' mozgu je zostavená z nameraných fMRI dát
- podstatou fMRI je zaznamenávanie zmien v prietoku krvi
- zaznamená sa vizuálna "mapa" aktivity rôznych častí mozgu
- skúmaní pacienti s Alzheimerovov chorobov alebo zdraví jedinci v mladom, či starom veku



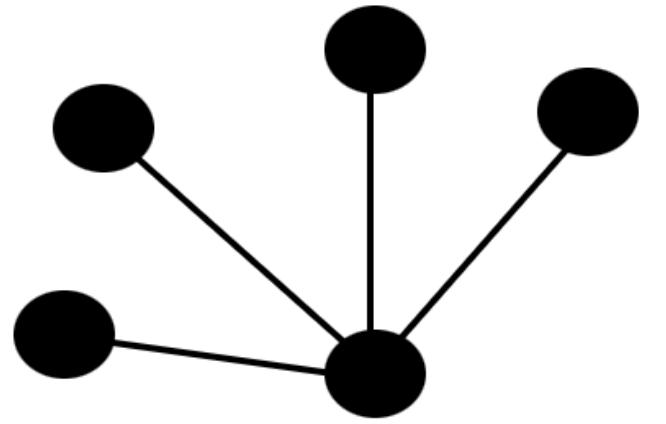
Definície

Ako merat štruktúru grafu?

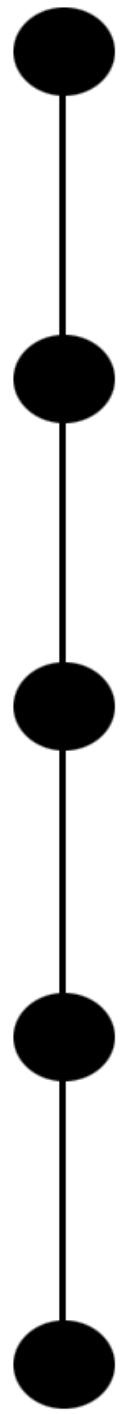


Definície

Priemerný stupeň vrchola



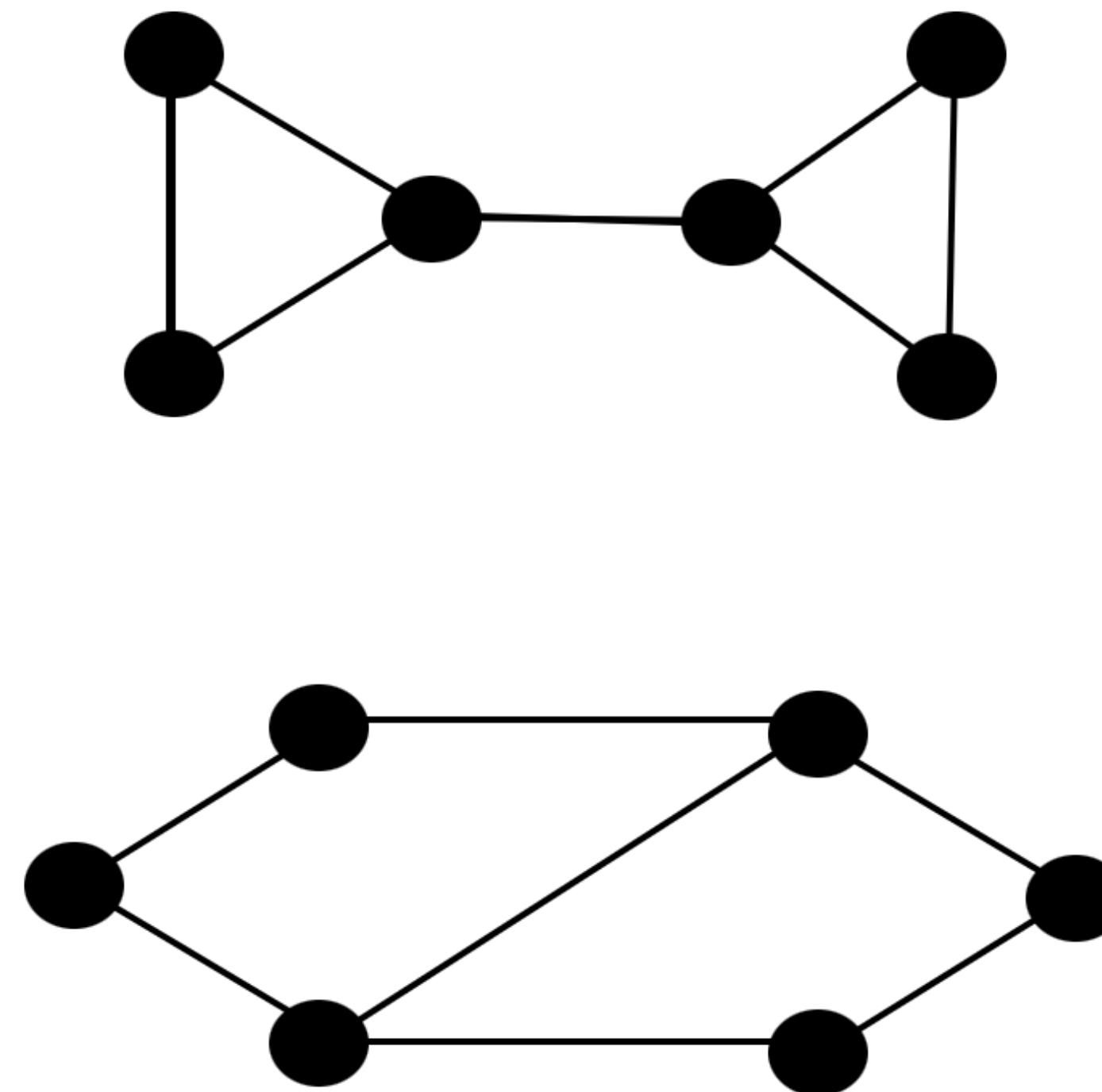
$$\bar{k} = \frac{8}{5}$$



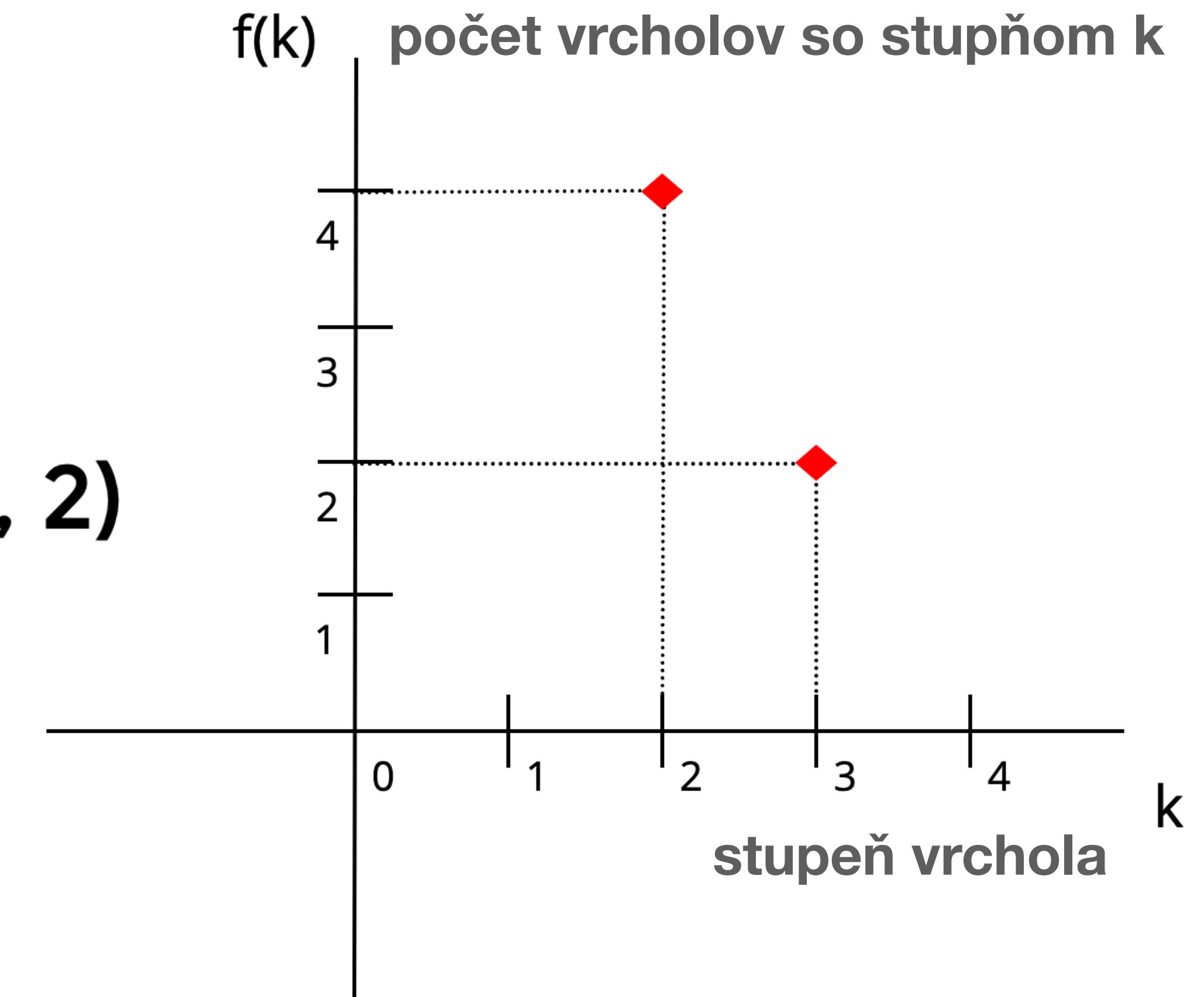
Definície

Distribúcia stupňa vrchola

(Postupnosť stupňov vrcholov)



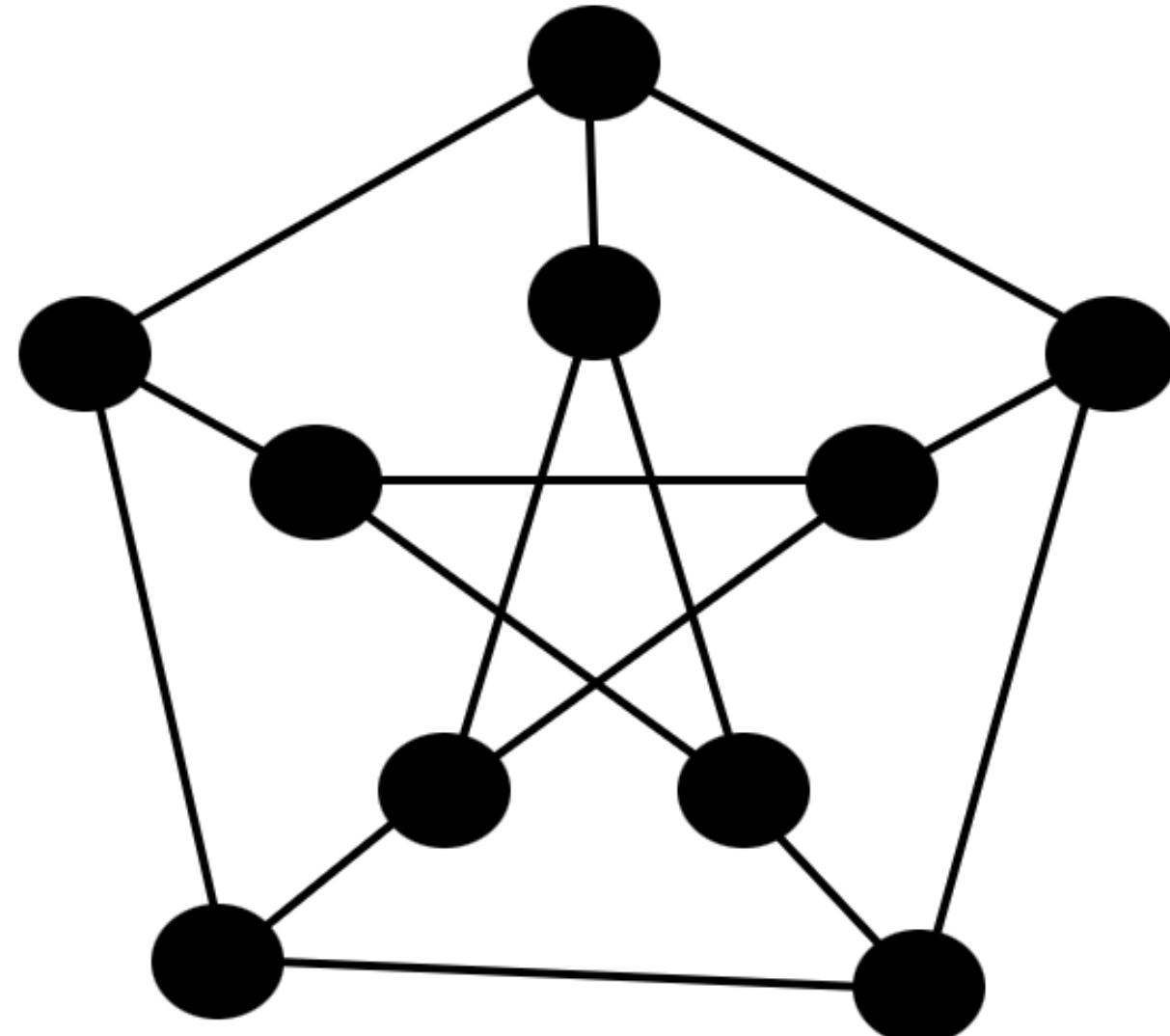
$(3, 3, 2, 2, 2, 2)$



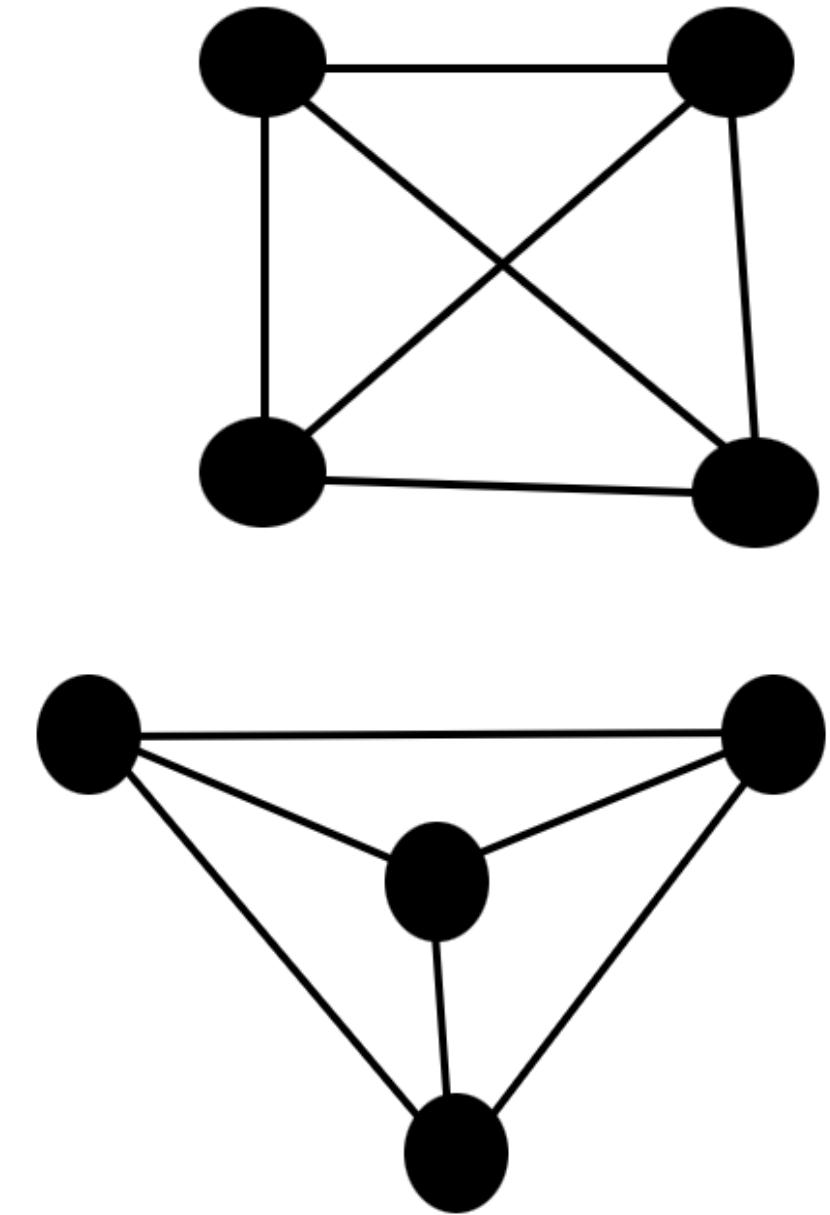
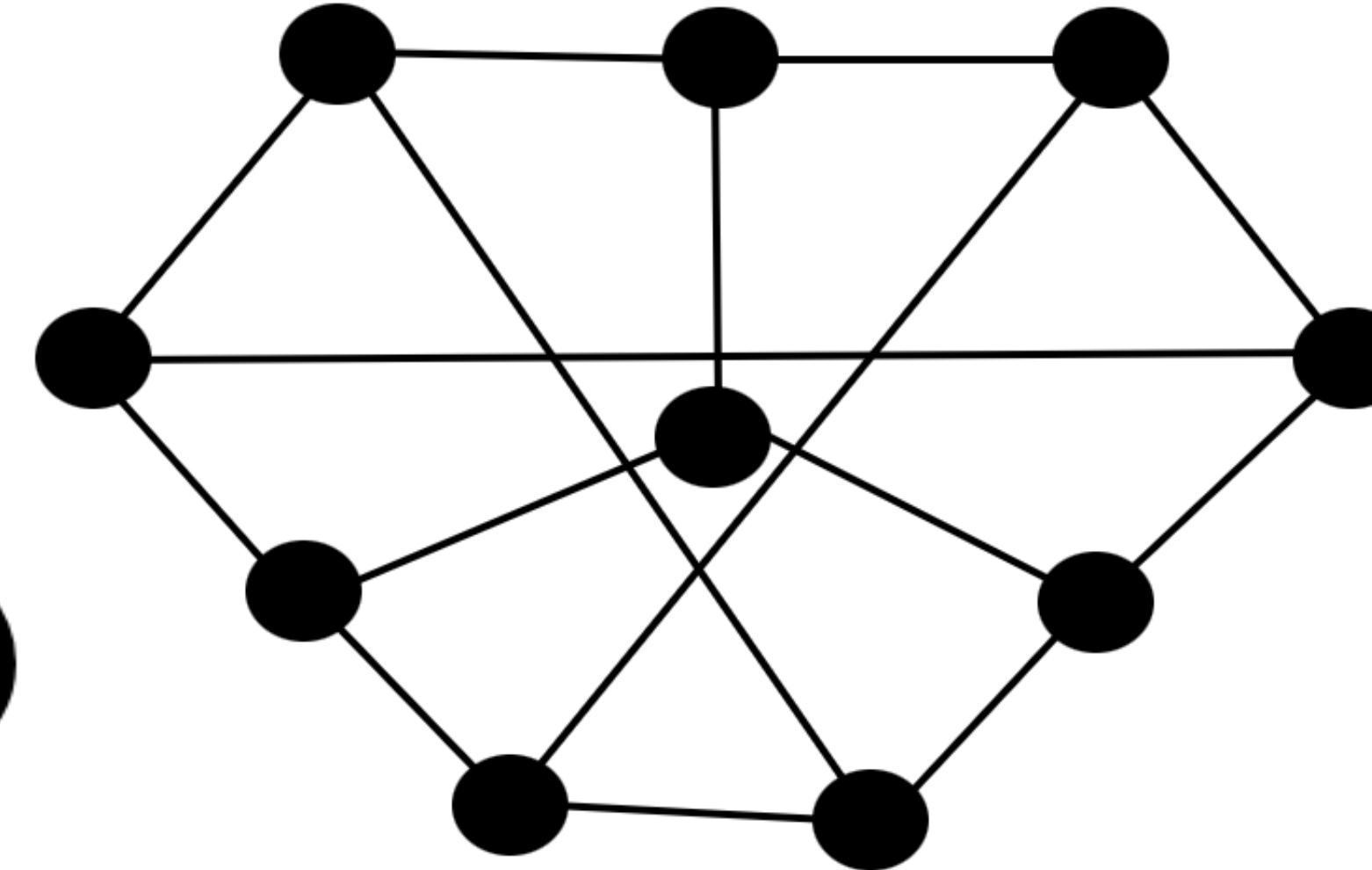
Definície

Izomorfizmus grafov

- bijekcia vrcholov 2 grafov G, H tak, že sa zachovajú hrany aj nehrany



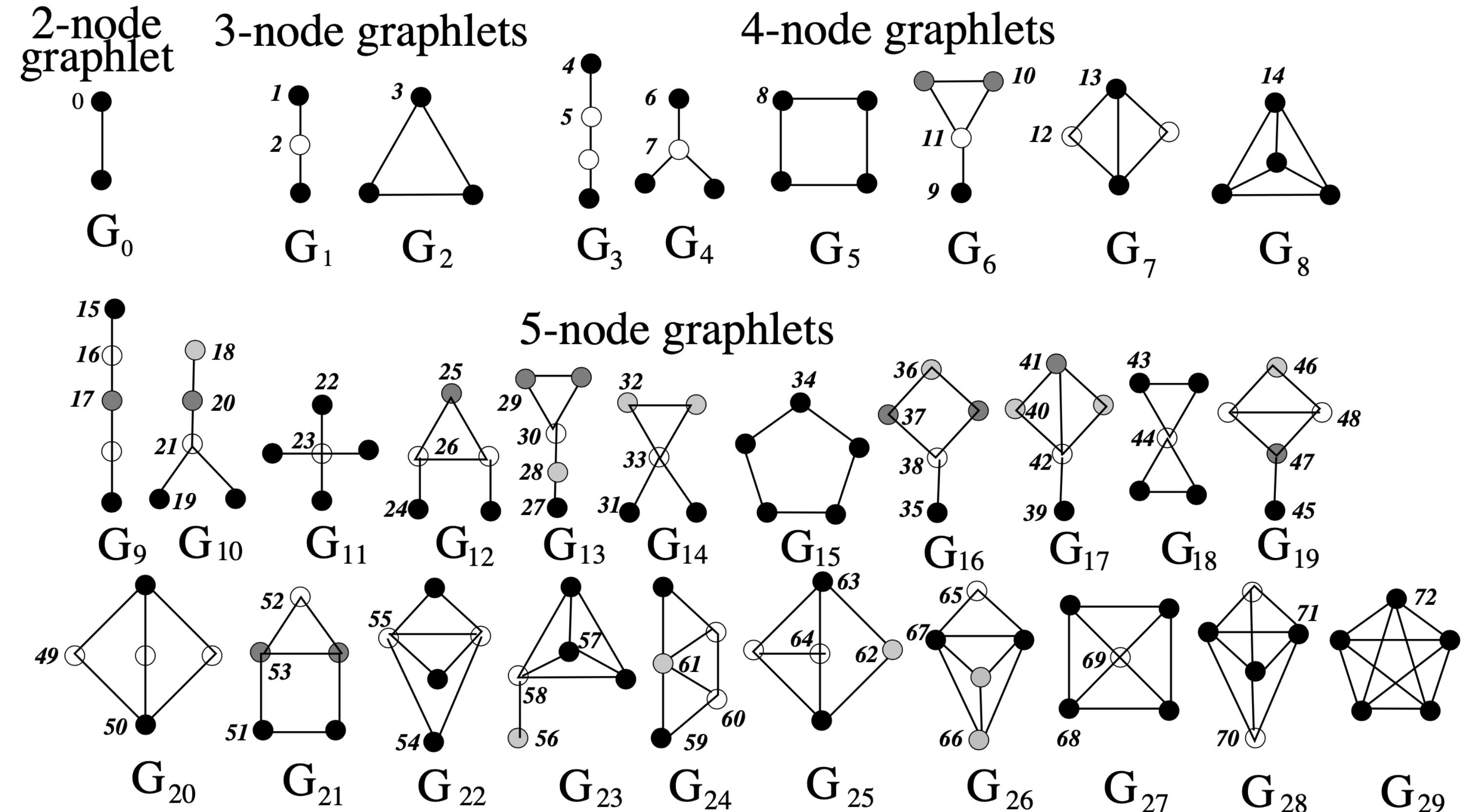
$O(n!)$



Definície

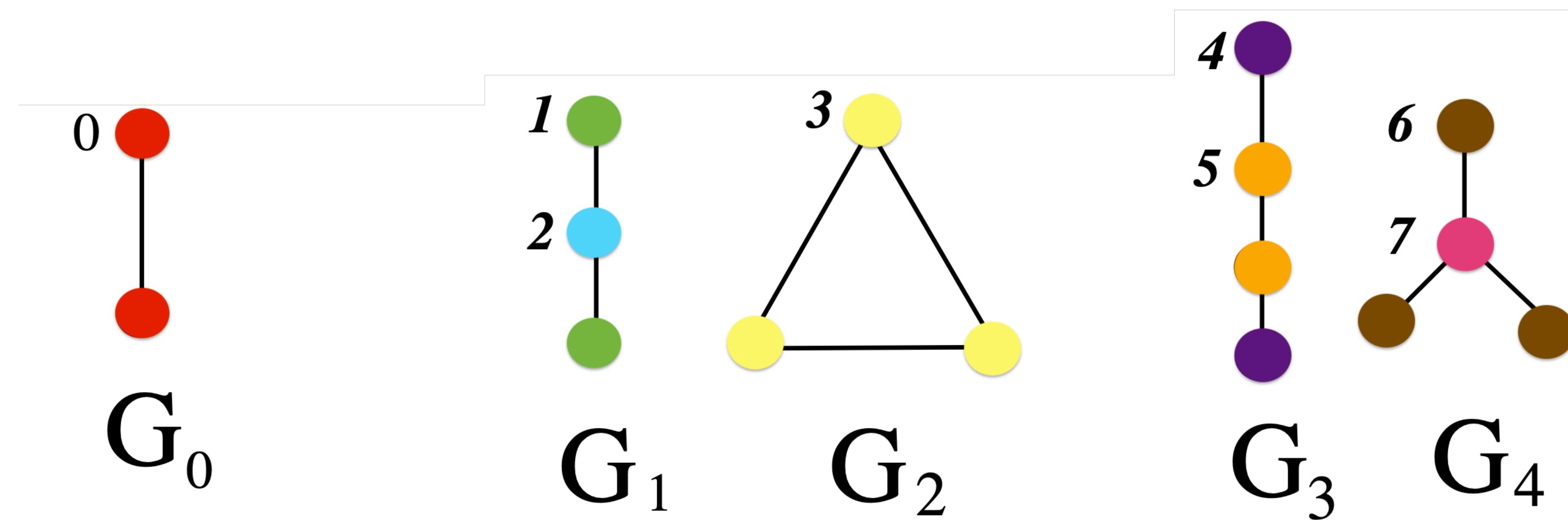
Graflet

- malý indukovaný podgraf veľkej siete
- indukovaný - ak vyberieme dané uzly pôvodného grafu, tak k nim musíme vybrať aj všetky pôvodné hrany a nehrany do vytvoreného podgrafa



Orbita

- množina uzlov v graflete, ktoré sa dajú na seba jednoznačne premietnuť a sú v topologicky rovnakých pozíciách



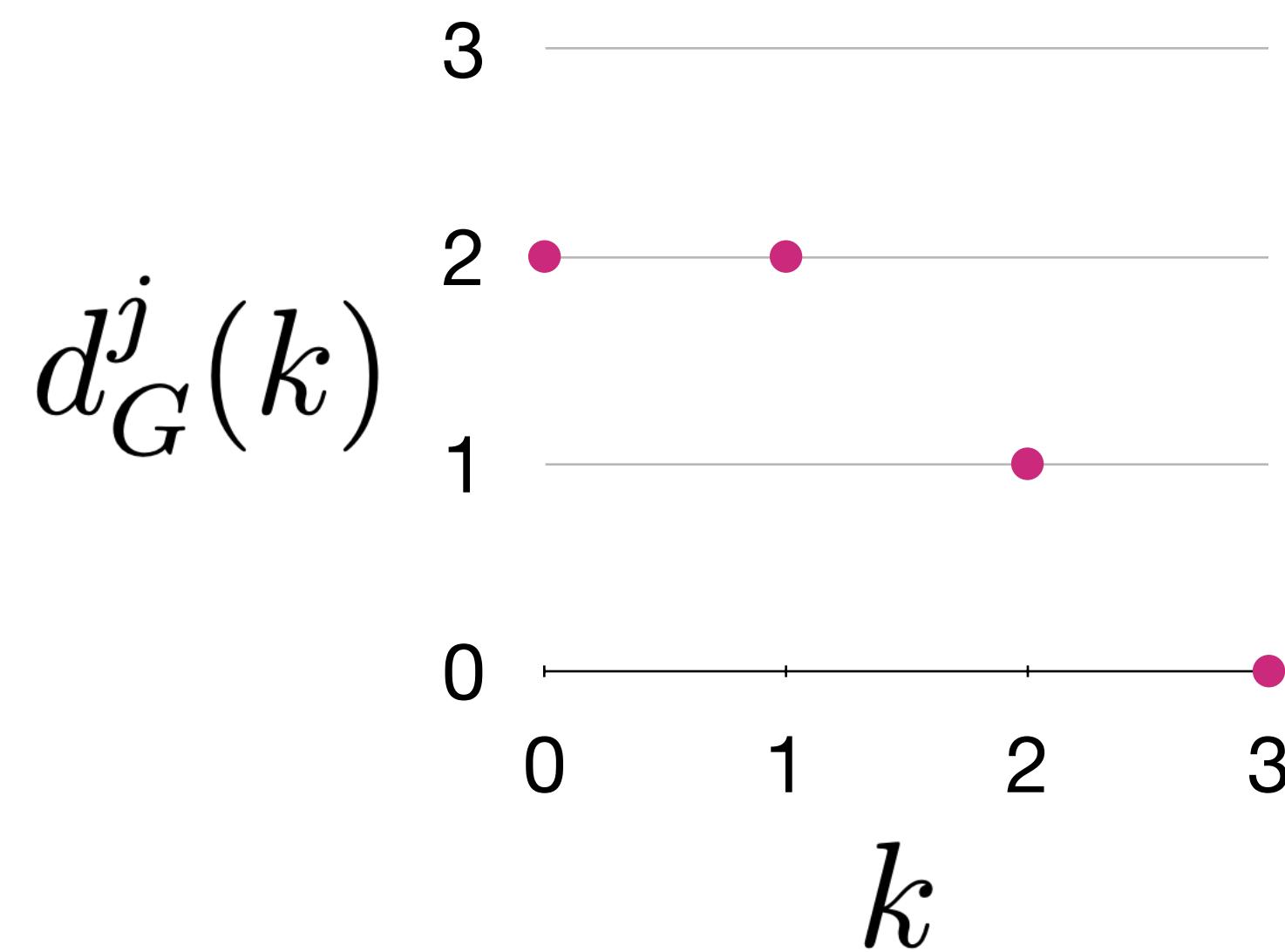
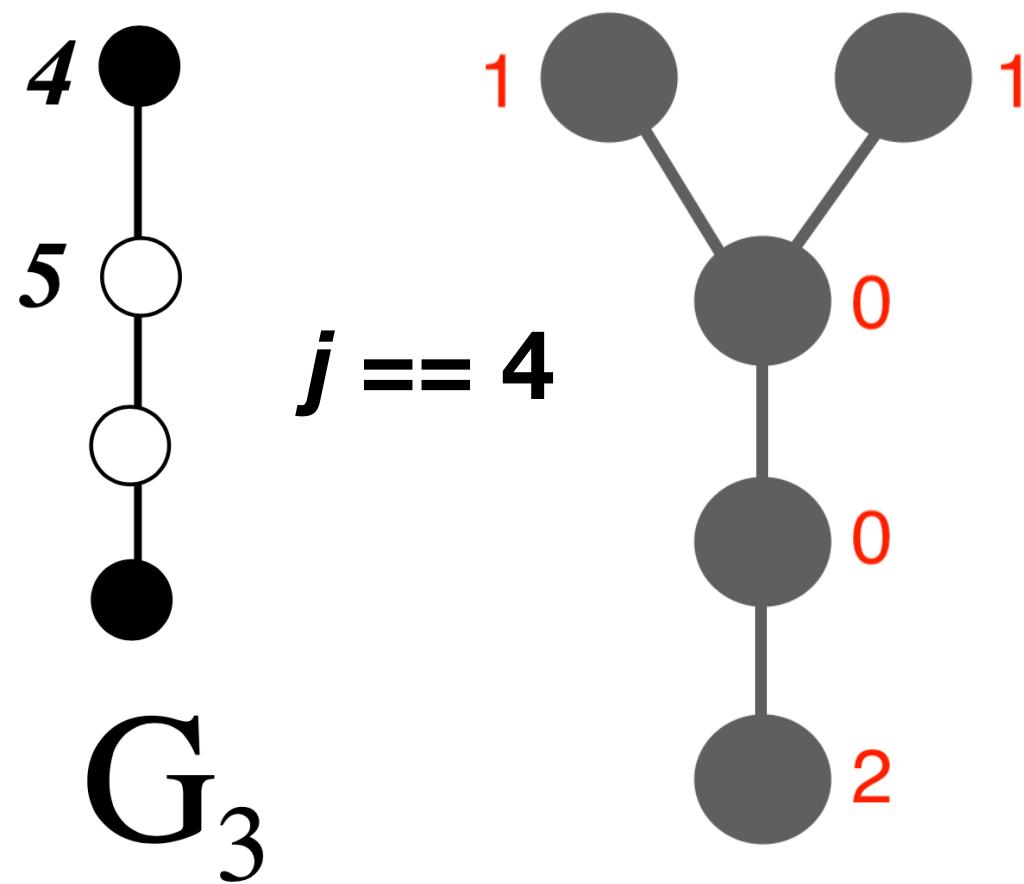
GDD

Graphlet Degree Distribution

- distribúcia stupňa obsiahnutých orbít grafletov
- pre danú orbitu j určíme počet všetkých uzlov v grafe, ktoré sú súčasťou daného grafletu tak, že sa nachádzajú v danej orbite j určitý počet krát

GDD

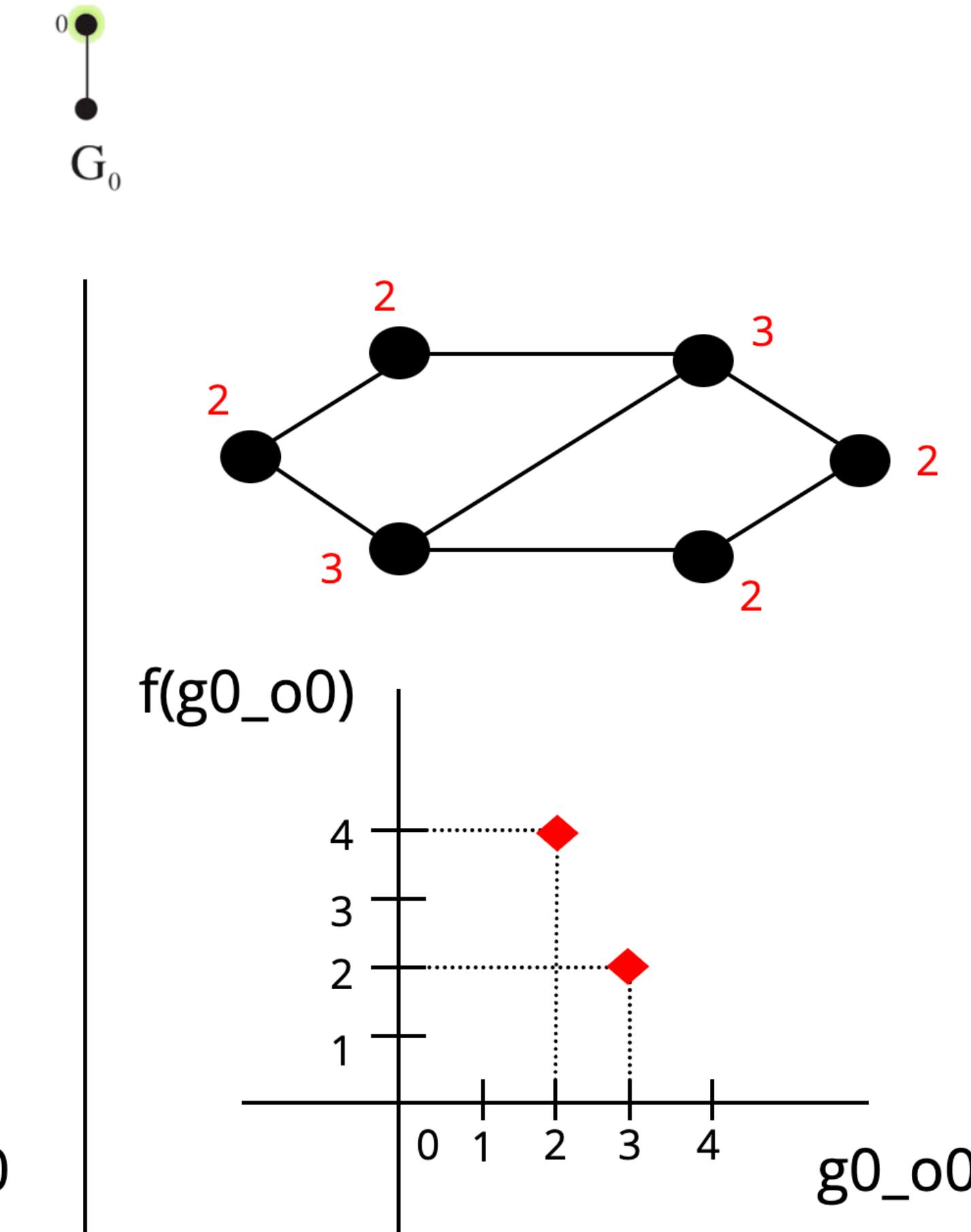
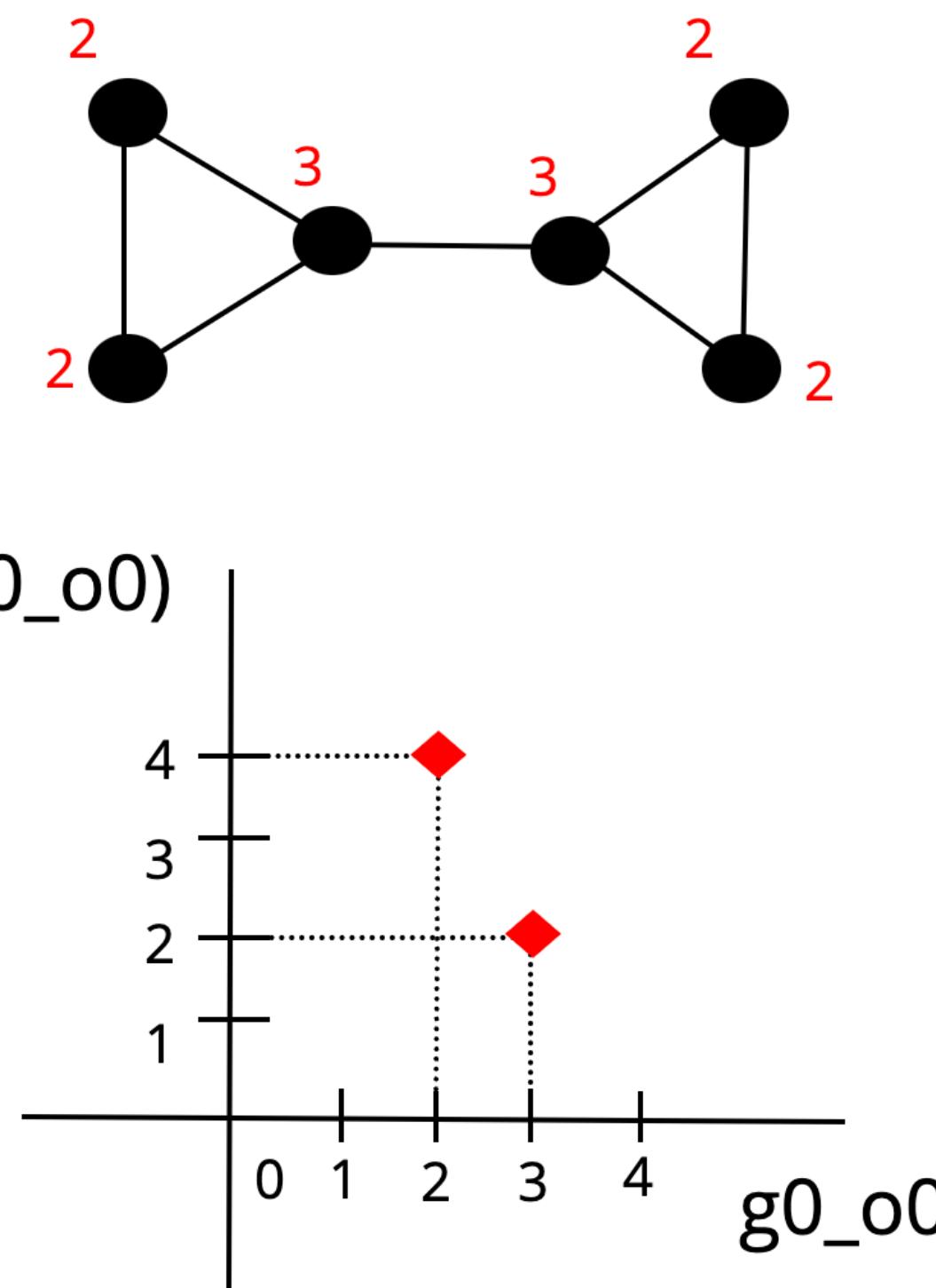
Graphlet Degree Distribution



k - počet kol'kokrát sa uzol vyskytuje v danej orbite j

$d_G^j(k)$ - počet všetkých uzlov v grafe, ktoré sa vyskytujú v orbite j daný počet k krát

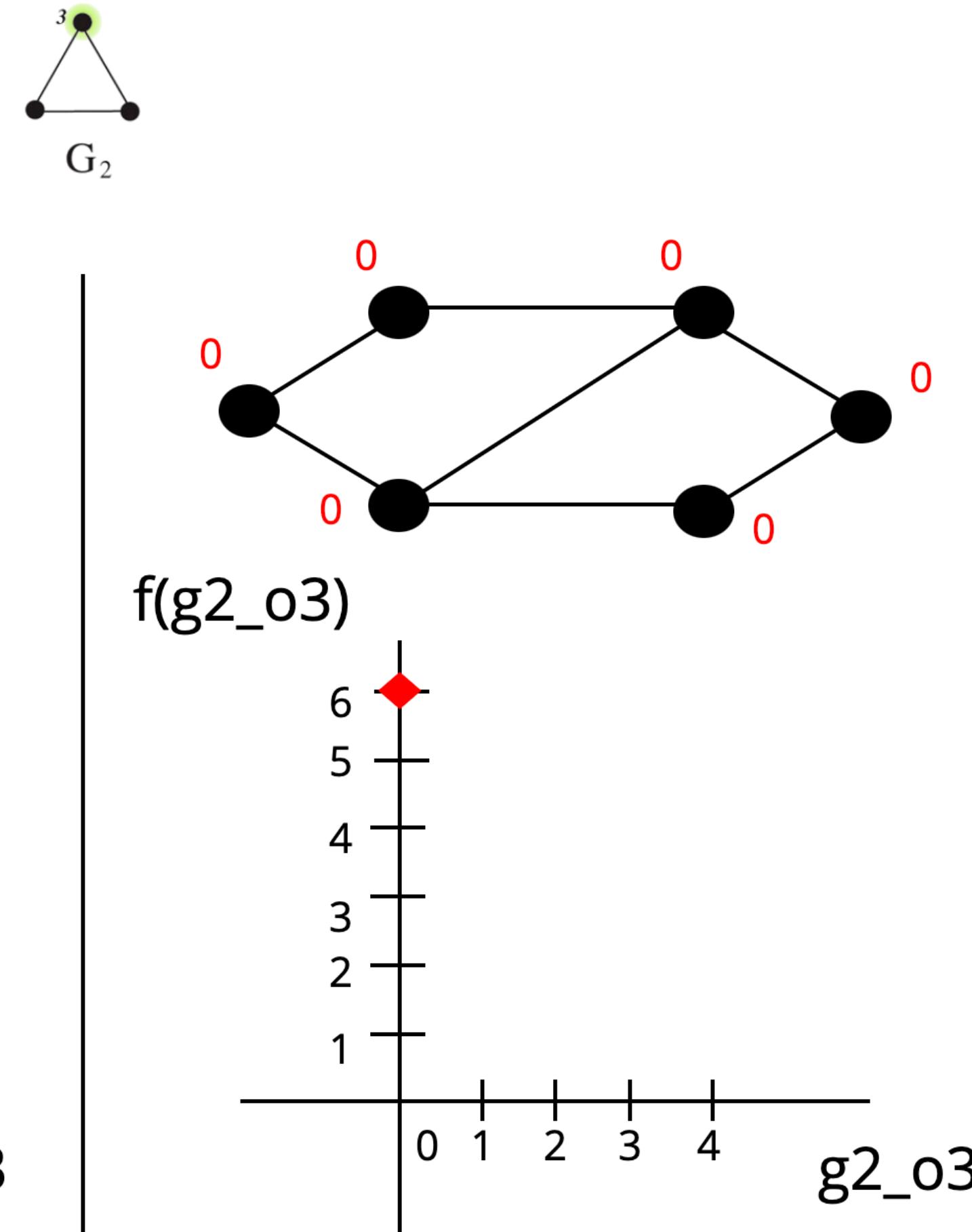
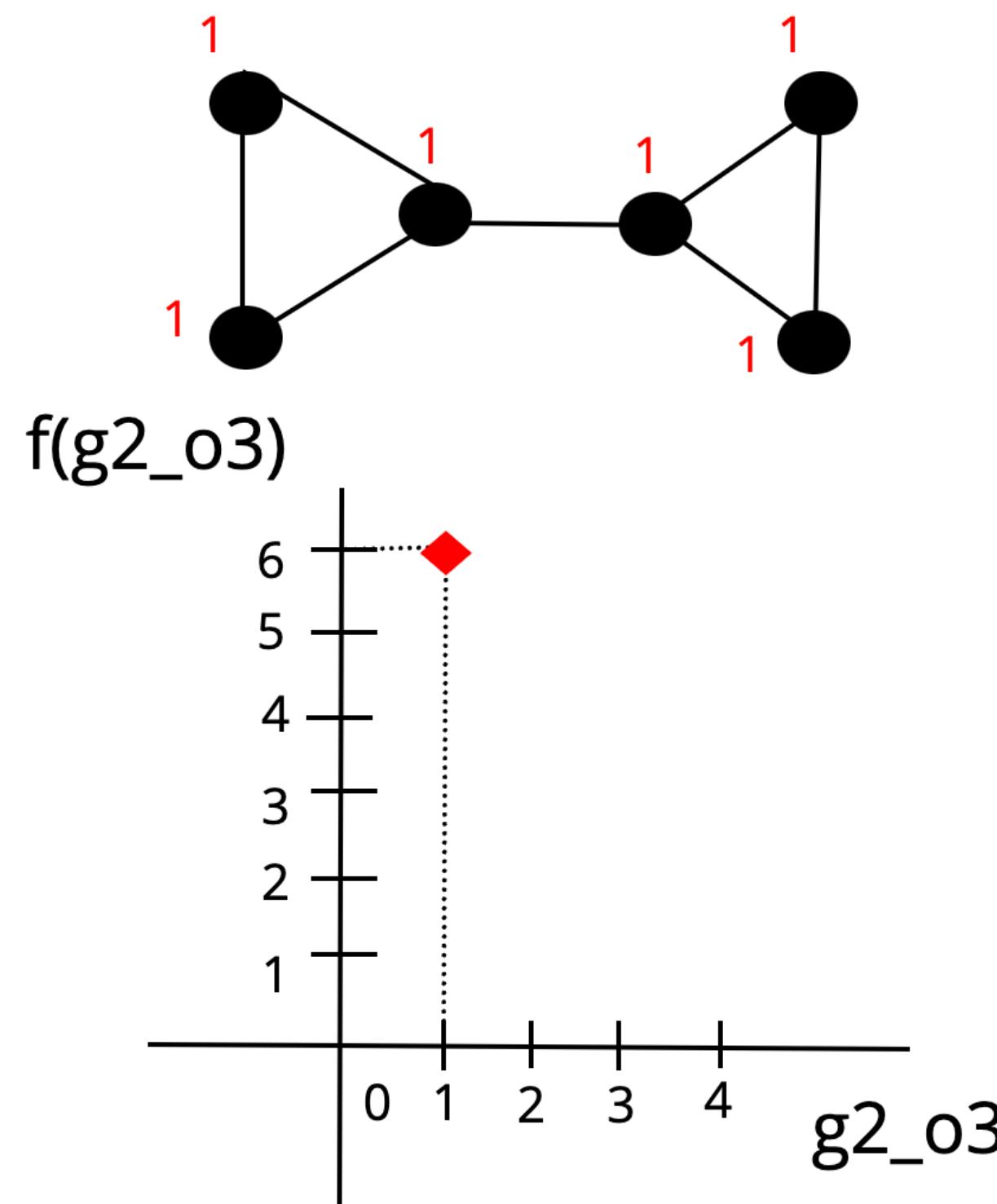
Porovnanie GDD orbity 0 grafetu 0



Definície

Nataša Pržulj (2007 / 2010)

Porovnanie GDD orbity 3 grafetu 2



Definície

Nataša Pržulj (2007 / 2010)

Súhlas GDD

GDD agreement

- škálované GDD
- aby sa odstránil vplyv vyšších stupňov GDD

$$S_G^j(k) = \frac{d_G^j(k)}{k}$$

k - počet kol'kokrát sa uzol vyskytuje v danej orbite j

$d_G^j(k)$ - počet všetkých uzlov v grafe, ktoré sa vyskytujú v orbite j daný počet k krát

Súhlas GDD

GDD agreement

- normalizované GDD
- vzhľadom na všetky hodnoty k výskytov uzla v orbite j , ktoré sa v grafe nachádzajú

$$T_G^j = \sum_{k=1}^{\infty} S_G^j(k)$$

$$N_G^j(k) = \frac{S_G^j(k)}{T_G^j}$$

Súhlas GDD

GDD agreement

- vzdialenosť normalizovaných GDD 2 grafov G, H pre orbitu j

$$D^j(G, H) = \left(\left(\sum_{k=1}^{\infty} [N_G^j(k) - N_H^j(k)]^2 \right)^{\frac{1}{2}} \right) \cdot \frac{1}{\sqrt{2}}$$

Súhlas GDD

GDD agreement

- súhlas GDD 2 grafov G, H pre orbitu j
 - hodnota 1 - pre identické GDD
 - hodnota 0 - ak ich GDD nemajú žiadny prienik

$$A^j(G, H) = 1 - D^j(G, H)$$

Súhlas GDD

GDD agreement

- aritmetický súhlas GDD 2 grafov G, H
- aritmetický priemer súhlasov GDD grafov cez všetky orbity j

$$A_{arith}(G, H) = \frac{1}{73} \sum_{j=0}^{72} A^j(G, H)$$

Súhlas GDD

GDD agreement

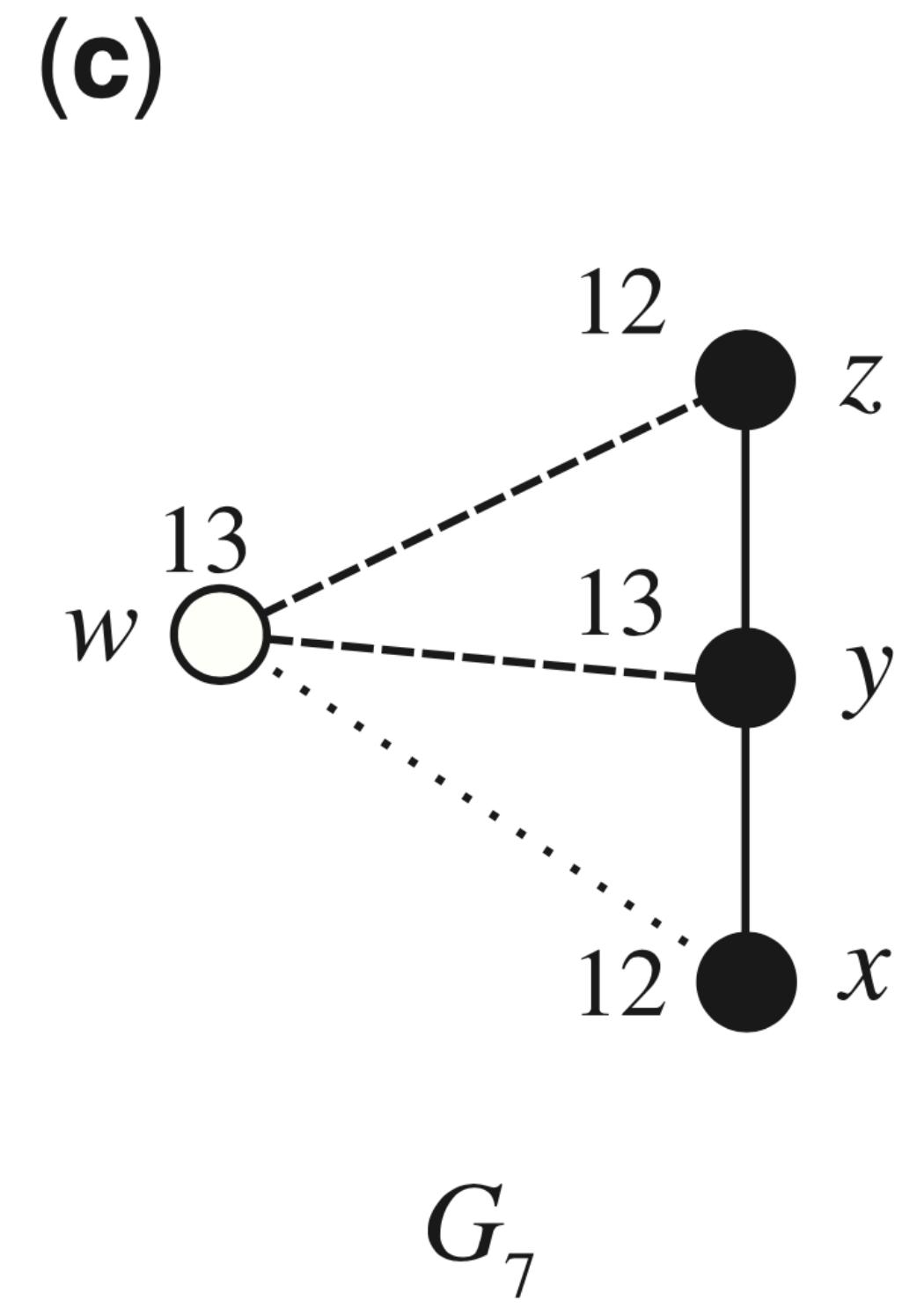
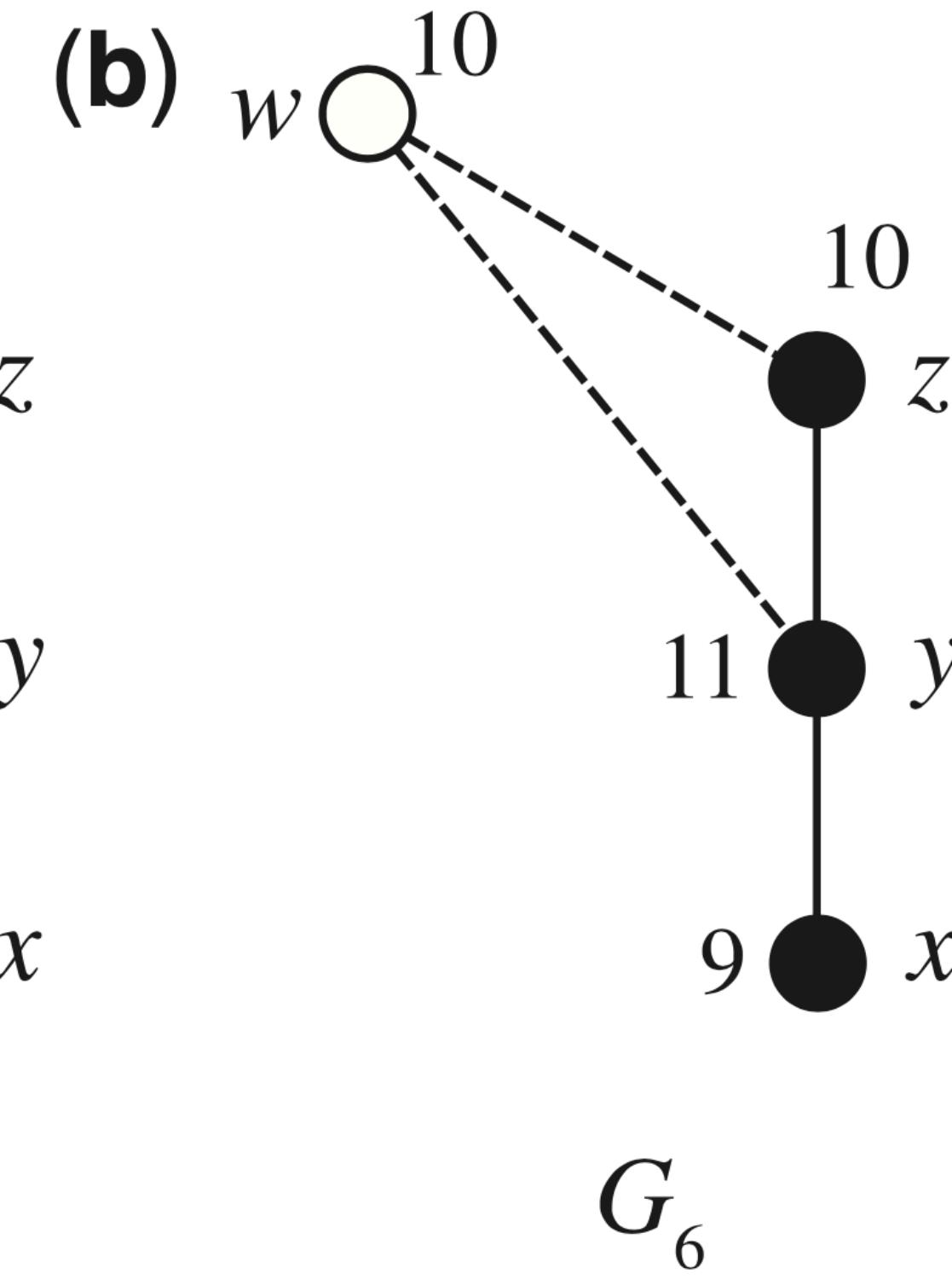
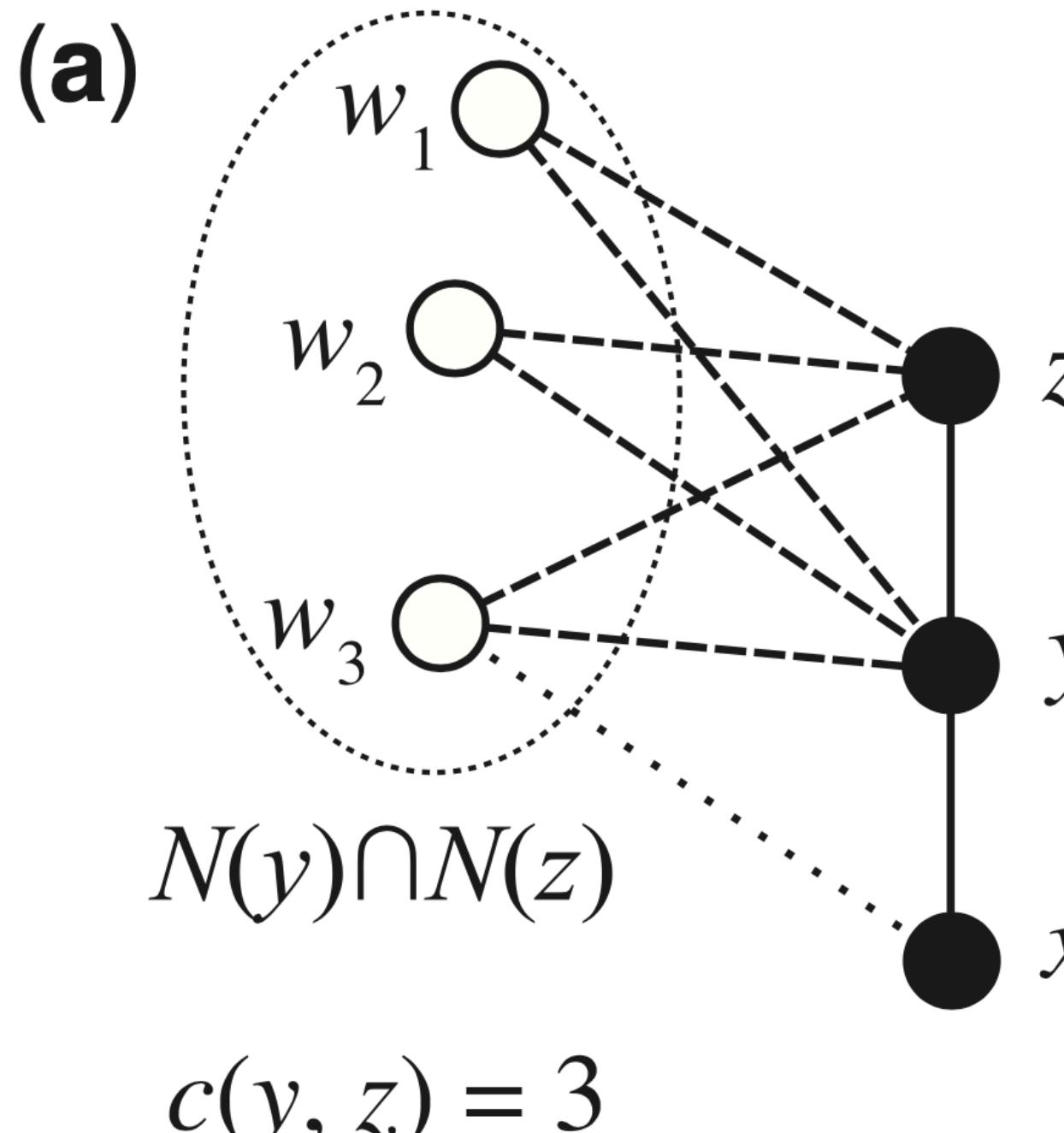
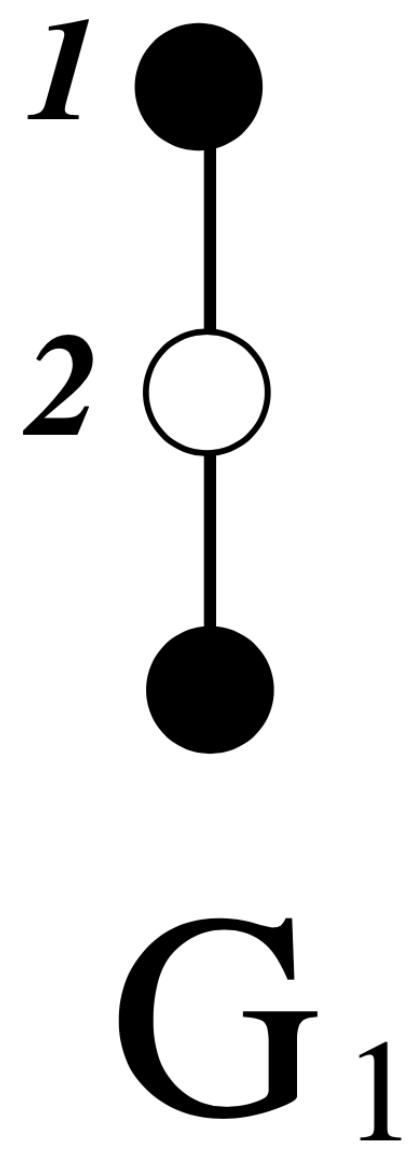
- geometrický súhlas GDD 2 grafov G, H
- geometrický priemer súhlasov GDD grafov cez všetky orbity j

$$A_{geo}(G, H) = \left(\prod_{j=0}^{72} A^j(G, H) \right)^{\frac{1}{73}}$$

ORCA

- kombinatorické riešenie počítania GDD
- sústavou lineárne nezávislých rovníc určíme počet výskytov každého uzla v grafe vo všetkých 73 orbitách
- pripojením 1 vrchola k n -vrcholovému grafletu vznikne $(n+1)$ -vrcholový graflet
- preskúmaním všetkých možností novovzniknutých grafletov dostaneme:
 - 10 rovníc pre 4 vrcholové grafety
 - 57 rovníc pre 5 vrcholové grafety

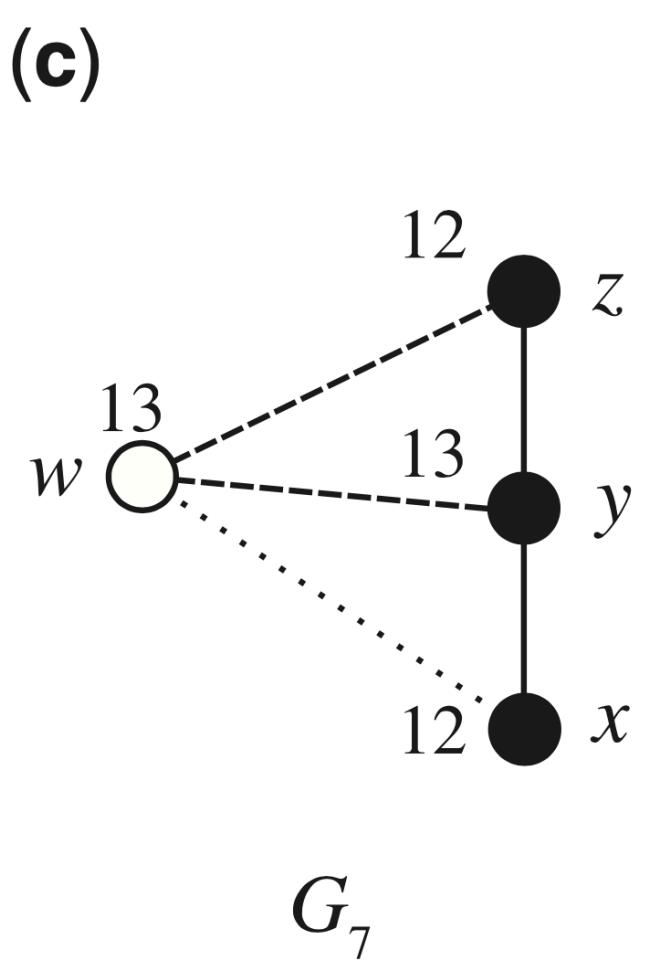
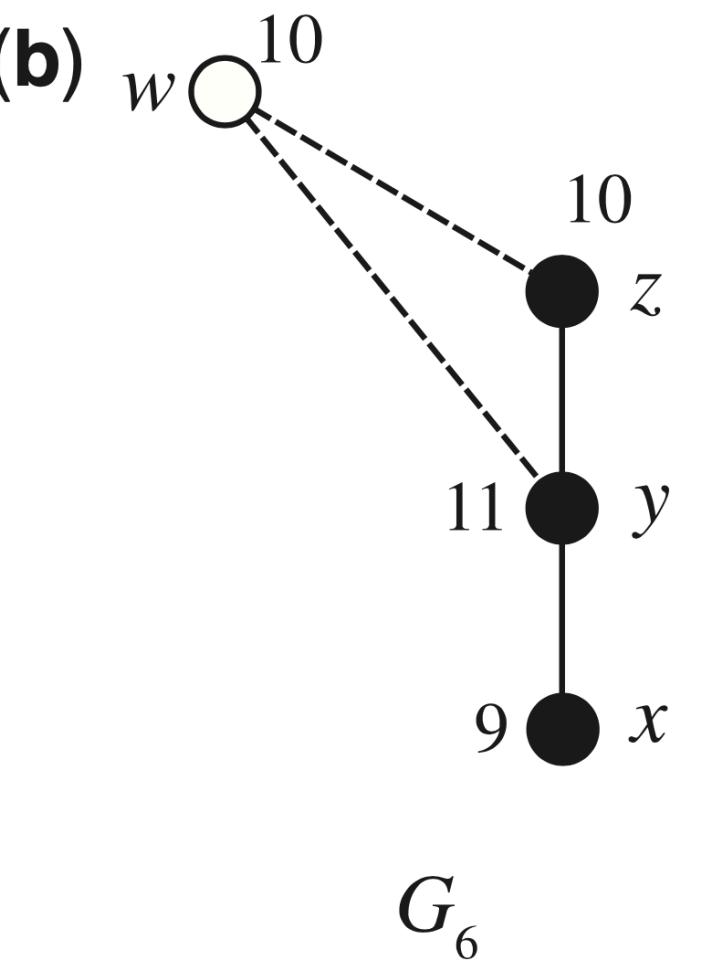
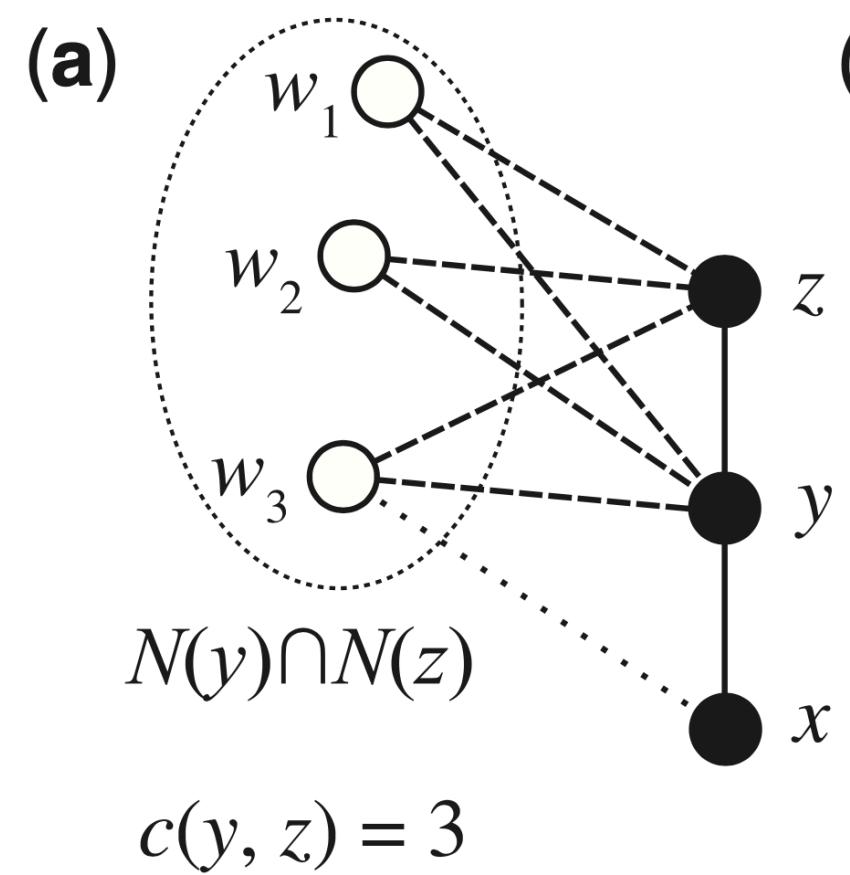
ORCA



Definície

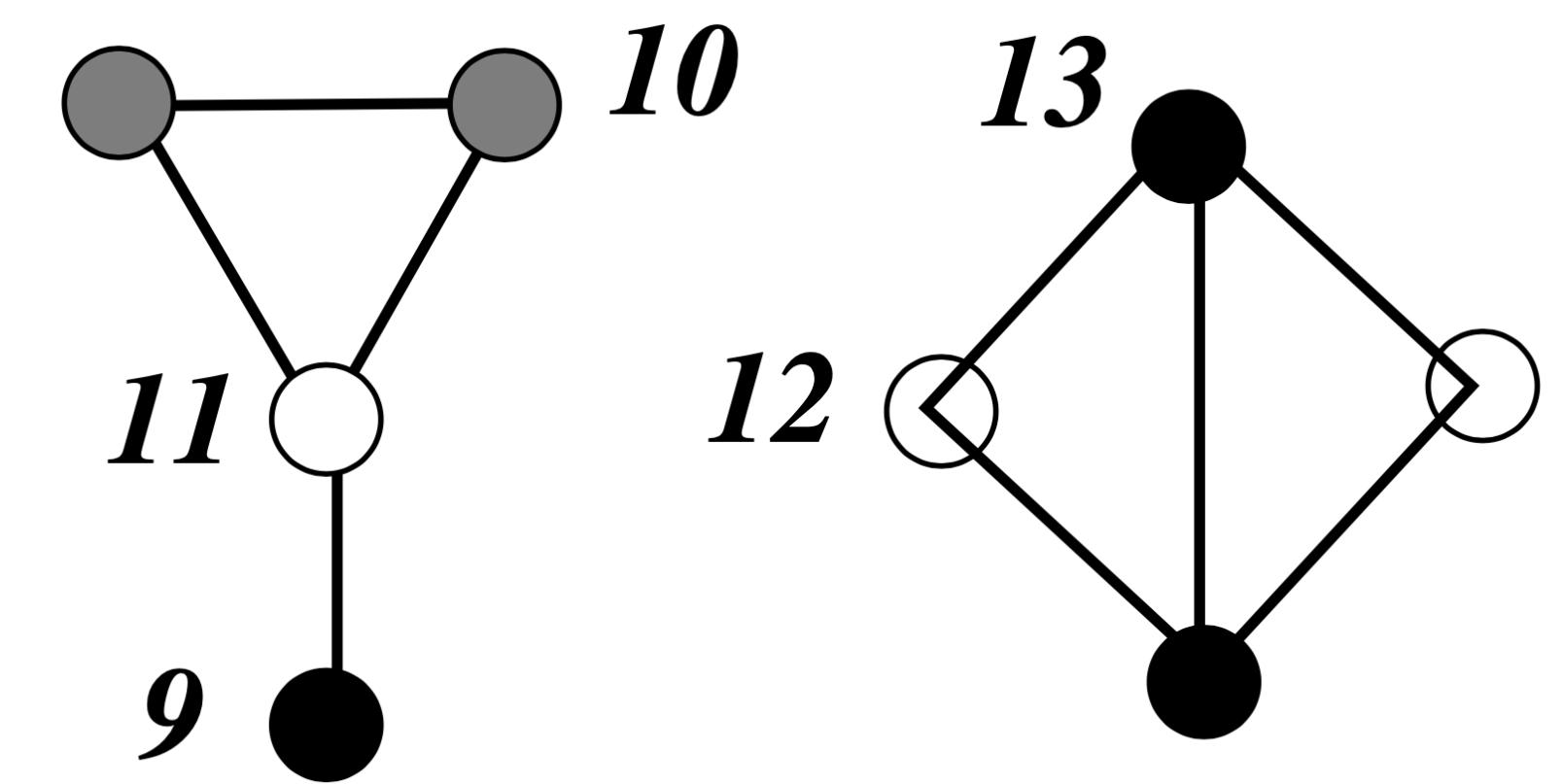
Tomaž Hočevar a Janez Demšar (2014)

ORCA



$$2o_9 + 2o_{12} = \sum_{\substack{y, z: x, z \in N(y) \\ G[\{x, y, z\}] \cong G_1}} c(y, z)$$

Definície



G_6

G_7

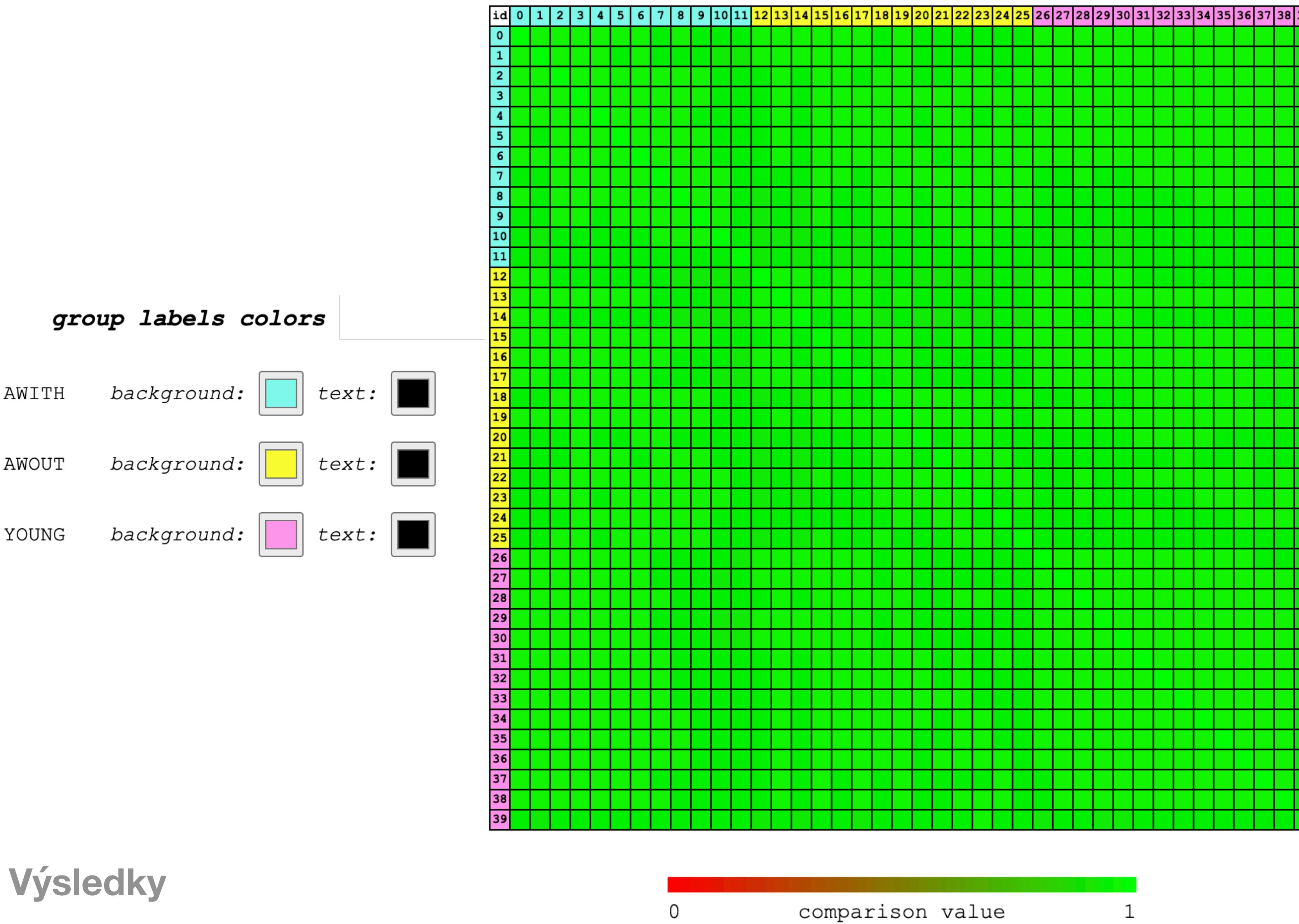
Tomaž Hočevar a Janez Demšar (2014)

Porovnanie súhlasov GDD funkčných sietí mozgu

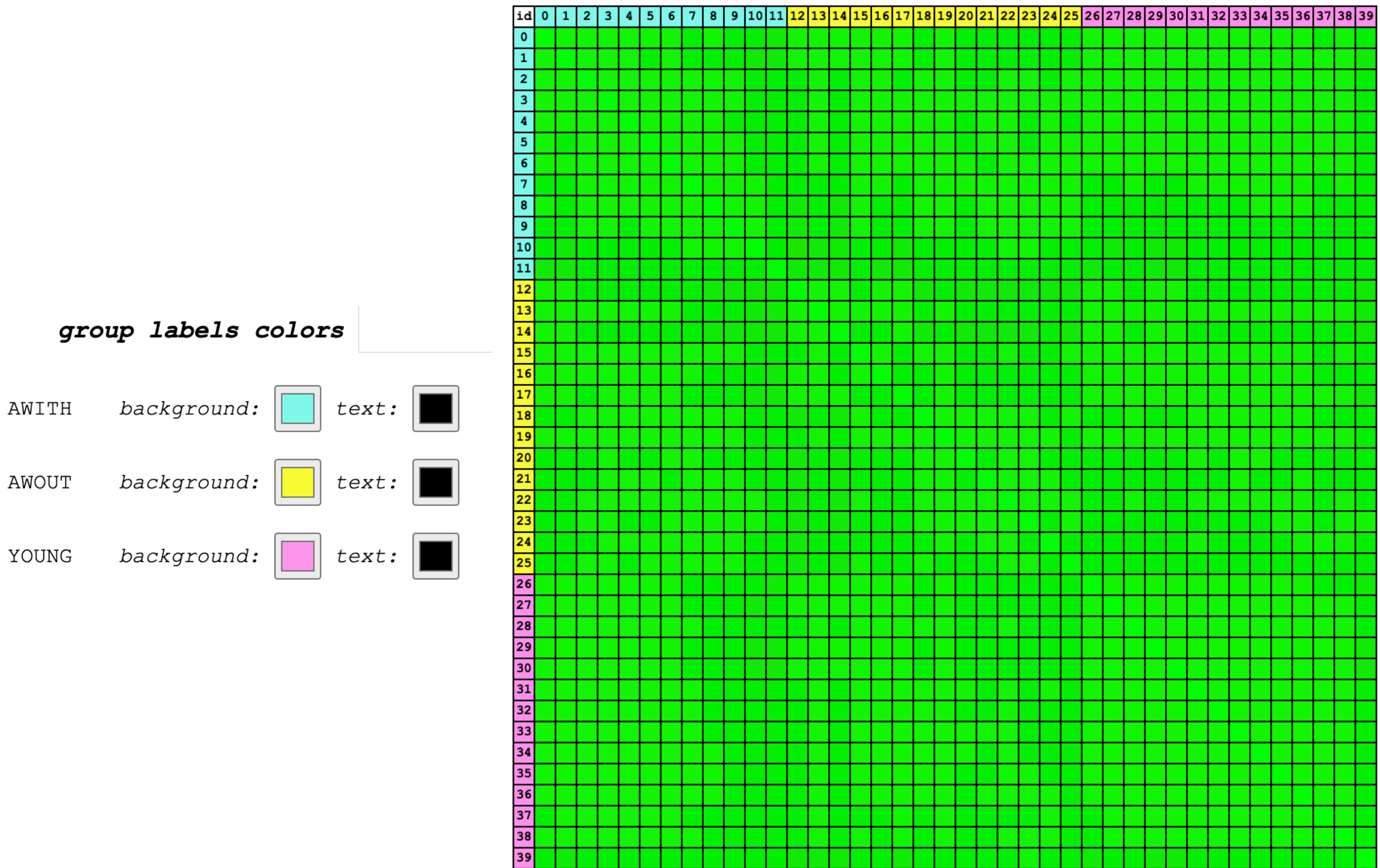
- skúmaní pacienti s Alzheimerovov chorobov alebo zdraví jedinci v mladom, či starom veku
- skupiny sa nedajú rozlíšiť kvôli zašumenosti dát
- súhlas GDD odhalí šum v dátach a vyhodnotí všetky grafy ako podobné

Výsledky

GDD agreement **arithmetic** comparison



GDD agreement **geometric** comparison

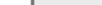
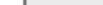


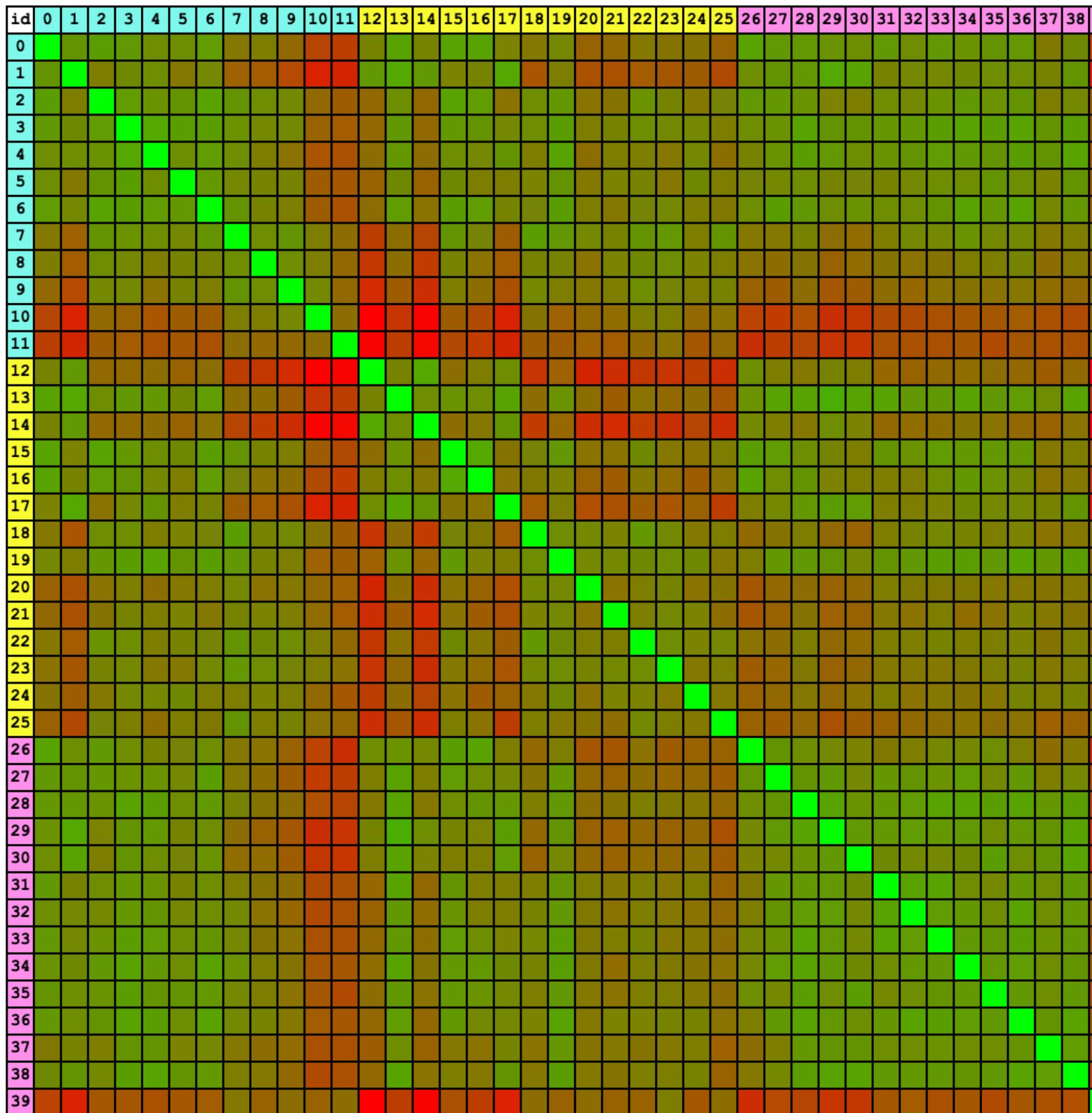
GDD agreement **arithmetic** comparison

group labels colors

AWITH *background:* *text:*

AWOUT *background:*  *text:* 

YOUNG *background:*  *text:* 



Výsledky



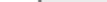
funkčné siete mozgu

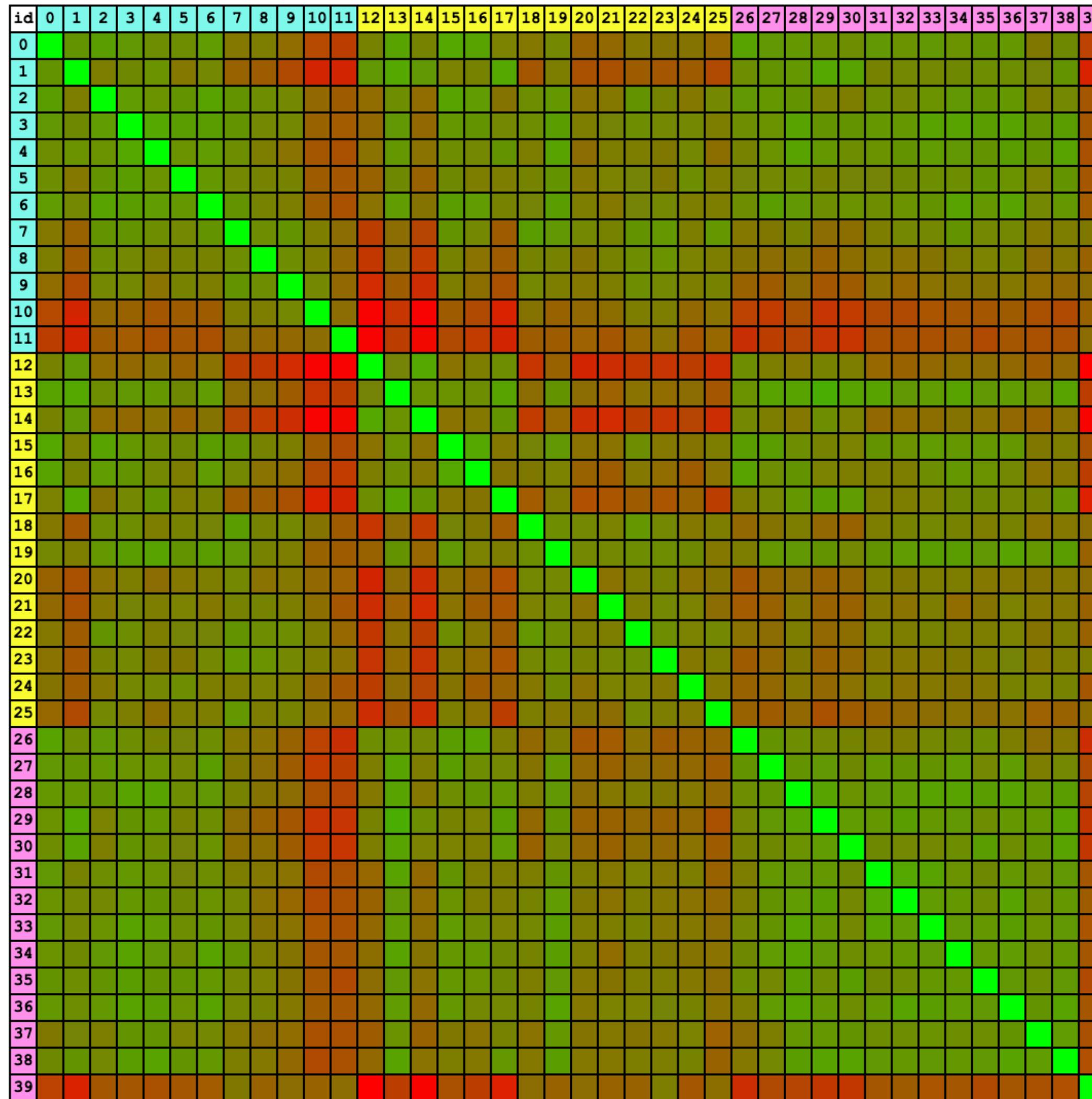
GDD agreement **geometric** comparison

group labels colors

AWITH *background:*  *text:* 

AWOUT *background:*  *text:* 

YOUNG *background:*  *text:* 



Výsledky



funkčné siete mozgu

Porovnanie súhlasov GDD funkčných sietí mozgu

- porovnané všetky grafy medzi sebou v skupine záujmu
- VERSUS porovnanie skupín A, B porovná každý graf zo skupiny A s každým grafom zo skupiny B
- vyhodnotili sme aritmetický priemer a štandardnú odchýlku nameraných hodnôt súhlasov GDD

$$\bar{x} = \frac{1}{N} \cdot \sum_{i=1}^N x_i$$

$$s_N = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^N (x_i - \bar{x})^2}$$

Výsledky

Porovnanie aritmetických súhlsov GDD funkčných sietí mozgu

group of graphs	arithmetic mean	standard deviation
awith	0.958951	0.009460
awout	0.955163	0.010079
young	0.963415	0.008977
awout \cup young	0.959025	0.009187
awith VERSUS awout	0.957246	0.010419
awith VERSUS young	0.959280	0.009668
awout VERSUS young	0.958780	0.007890

Výsledky

Porovnanie geometrických súhlsov GDD funkčných sietí mozgu

group of graphs	arithmetic mean	standard deviation
awith	0.958696	0.009565
awout	0.954893	0.010151
young	0.963208	0.009084
awout \cup young	0.958783	0.009277
awith VERSUS awout	0.956986	0.010510
awith VERSUS young	0.959035	0.009771
awout VERSUS young	0.958535	0.007979

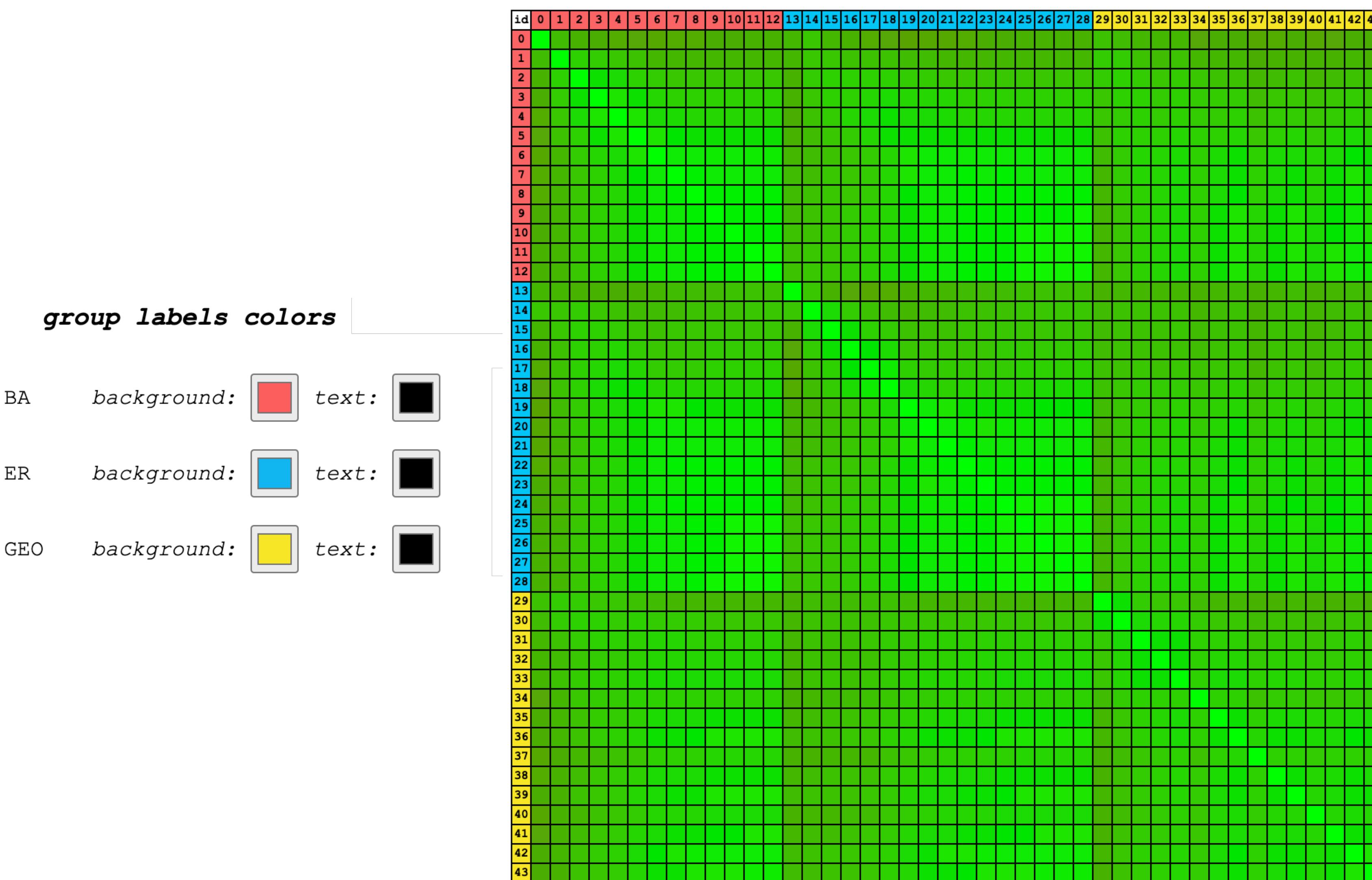
Výsledky

Porovnanie súhlsov GDD umelo vytvorených sietí

- BA, ER, GEO siete
- dáta z internetu s rôznym počtom uzlov a hrán
- súhlas GDD nedokáže rozlíšiť skupiny kvôli rôznemu počtu uzlov a hrán

Výsledky

GDD agreement **arithmetic** comparison



Výsledky

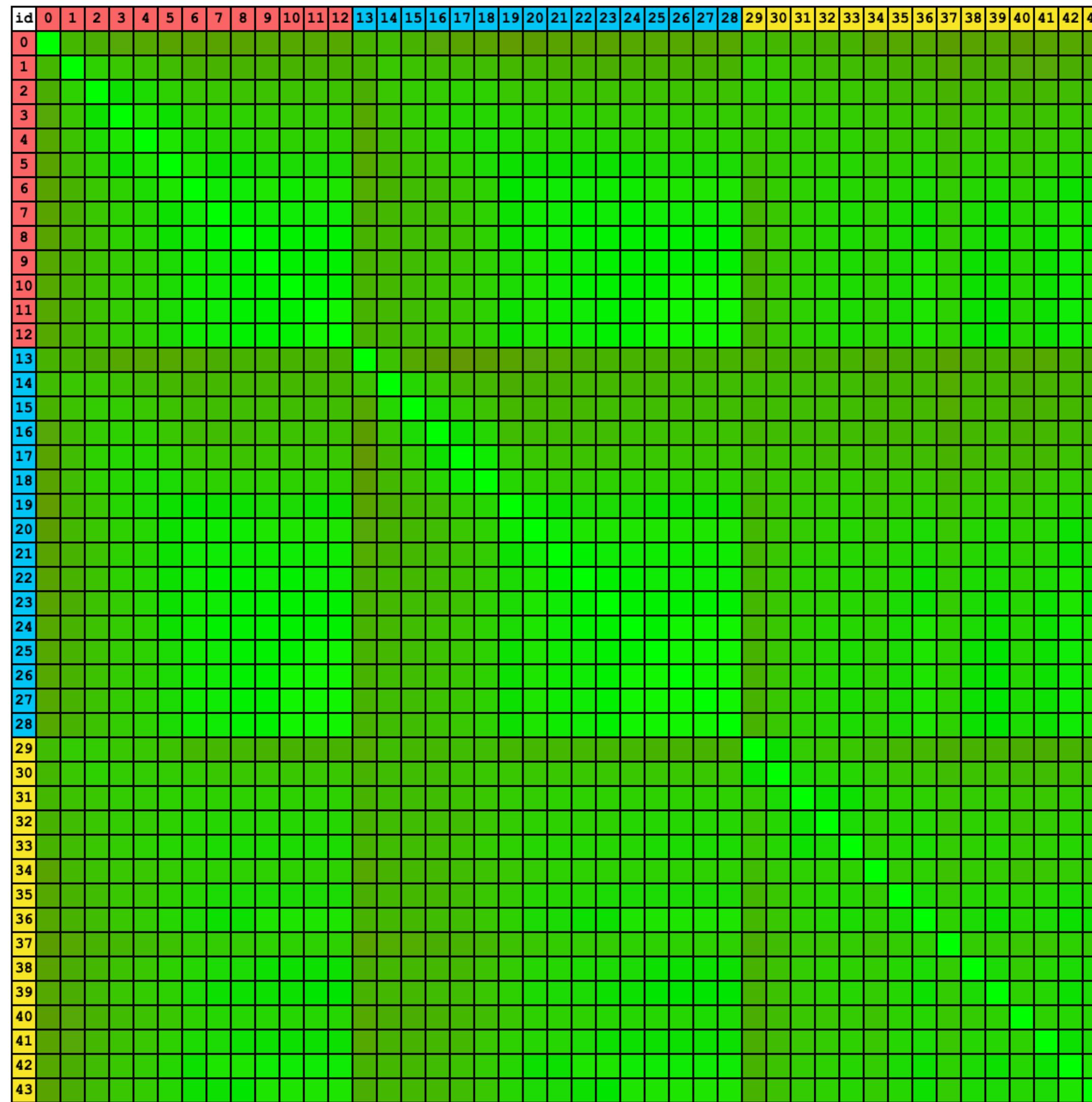


umelo vytvorené siete

GDD agreement **geometric** comparison

group labels colors

BA	<i>background:</i>		<i>text:</i>	
ER	<i>background:</i>		<i>text:</i>	
GEO	<i>background:</i>		<i>text:</i>	



Výsledky

umelo vytvorené siete

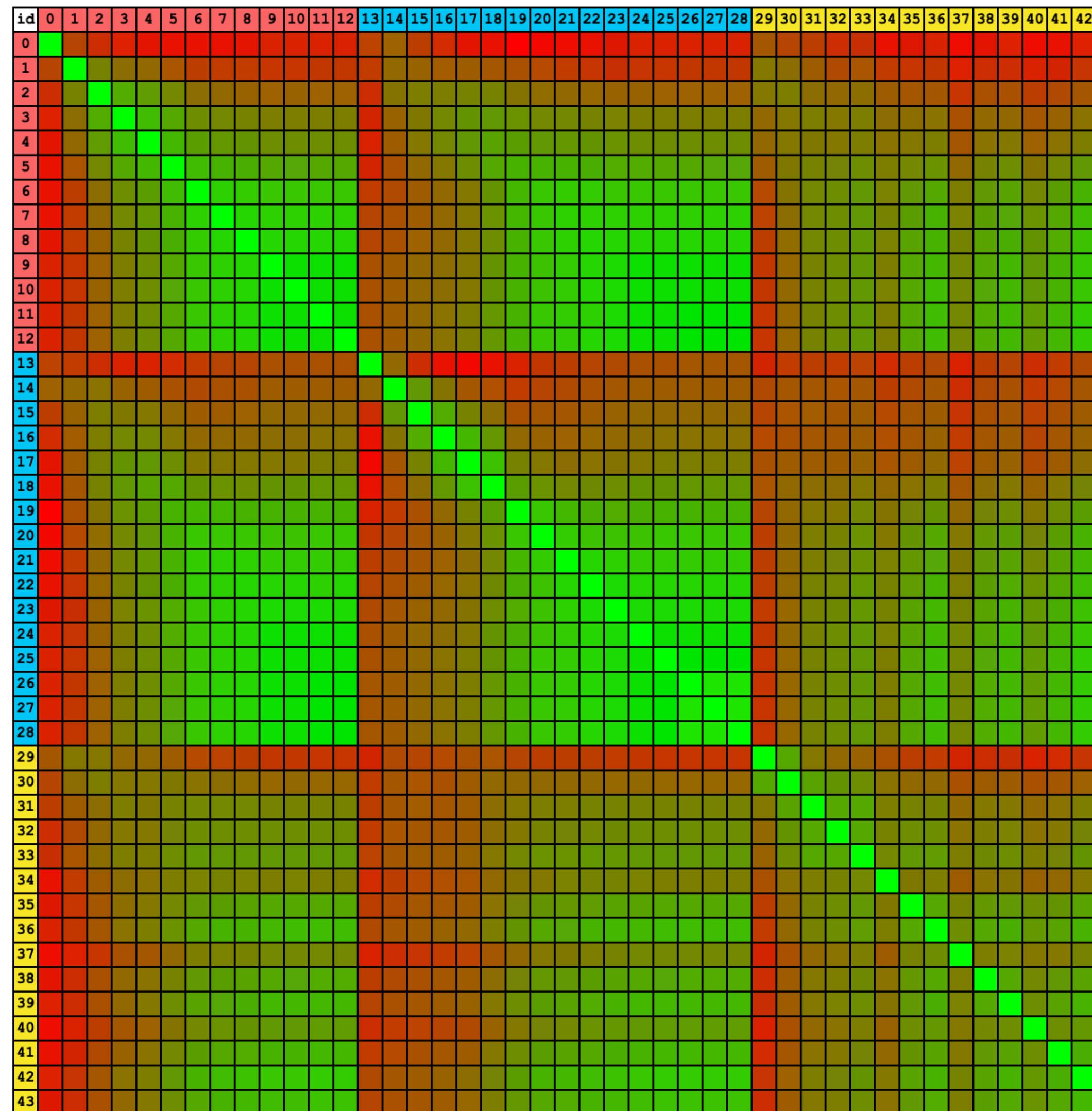
GDD agreement **arithmetic** comparison

group labels colors

BA *background:*  *text:* 

ER *background:*  *text:* 

GEO *background:*  *text:* 



Výsledky



umelo vytvorené siete

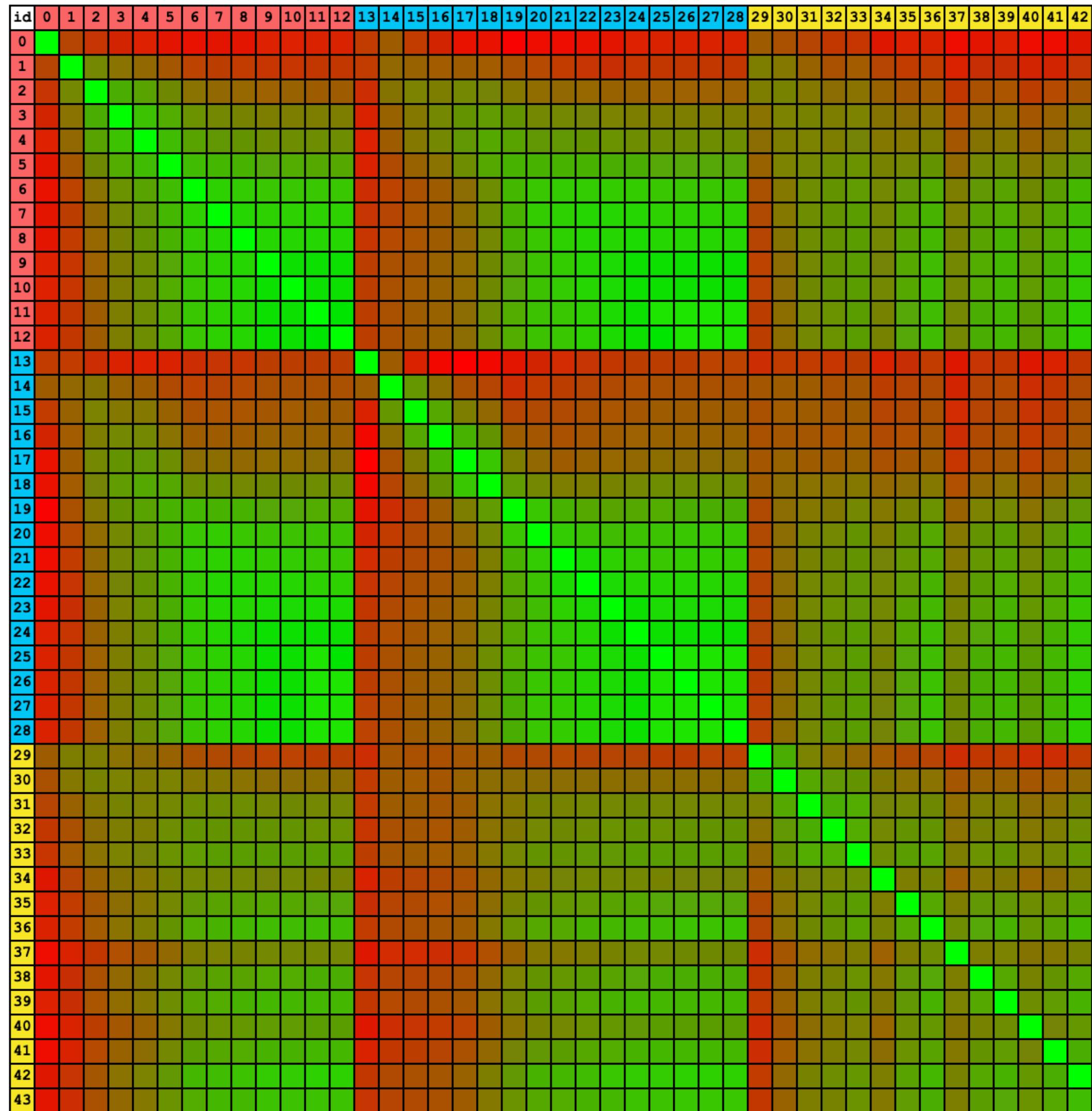
GDD agreement **geometric** comparison

group labels colors

BA *background:*  *text:* 

ER *background:*  *text:* 

GEO *background:*  *text:* 



Výsledky

0.614191 comparison value 1

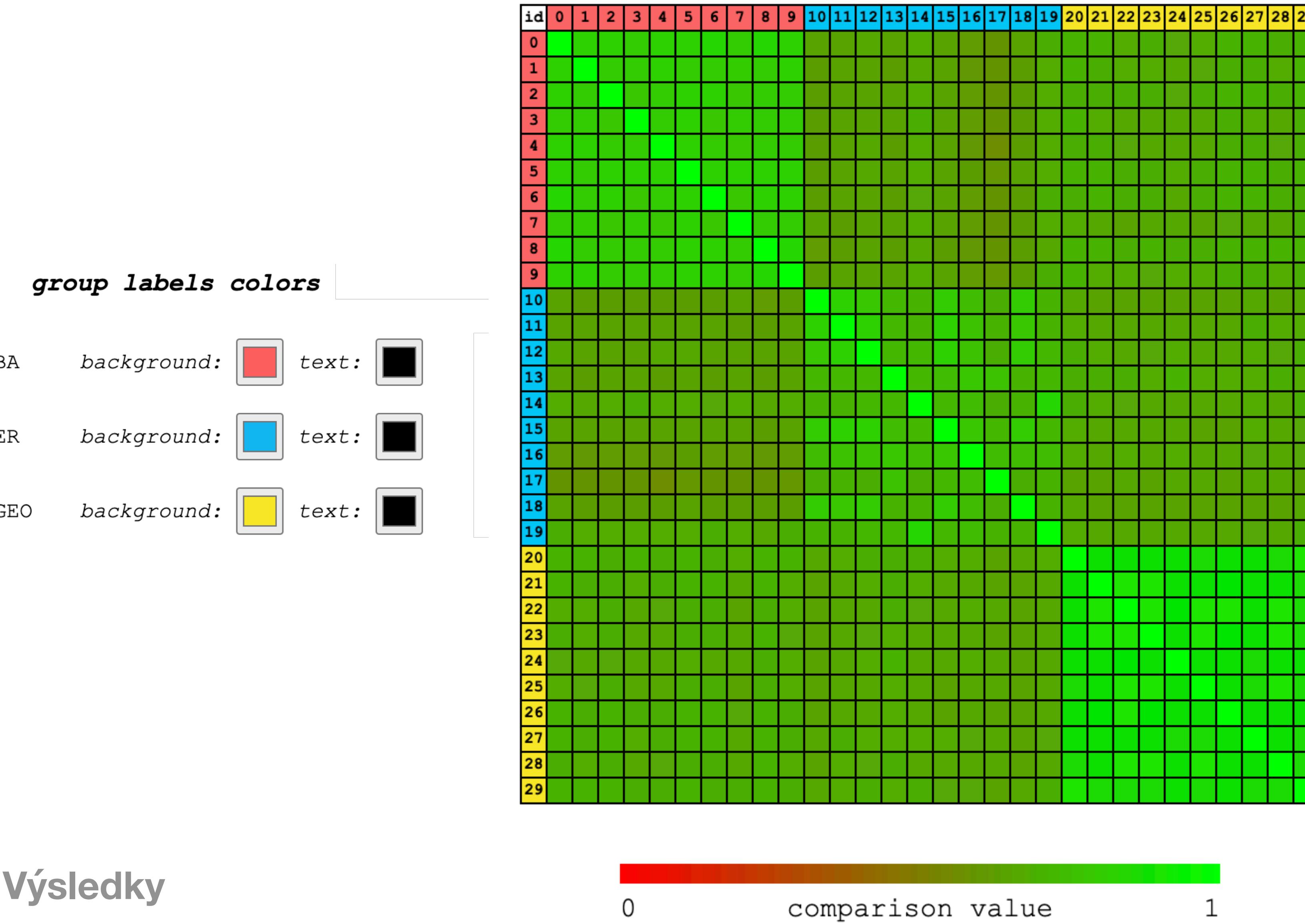
umelo vytvorené siete

Porovnanie súhlasov GDD umelo vytvorených sietí

- BA, ER, GEO siete
- vygenerované dáta s približne rovnakým počtom uzlov a hrán
- súhlas GDD rozlíši skupiny grafov s triviálne odlišnou štruktúrou
- úspešne sme overili súhlas GDD ako vhodnú mieru na porovnávanie štruktúry grafov

Výsledky

GDD agreement **arithmetic** comparison



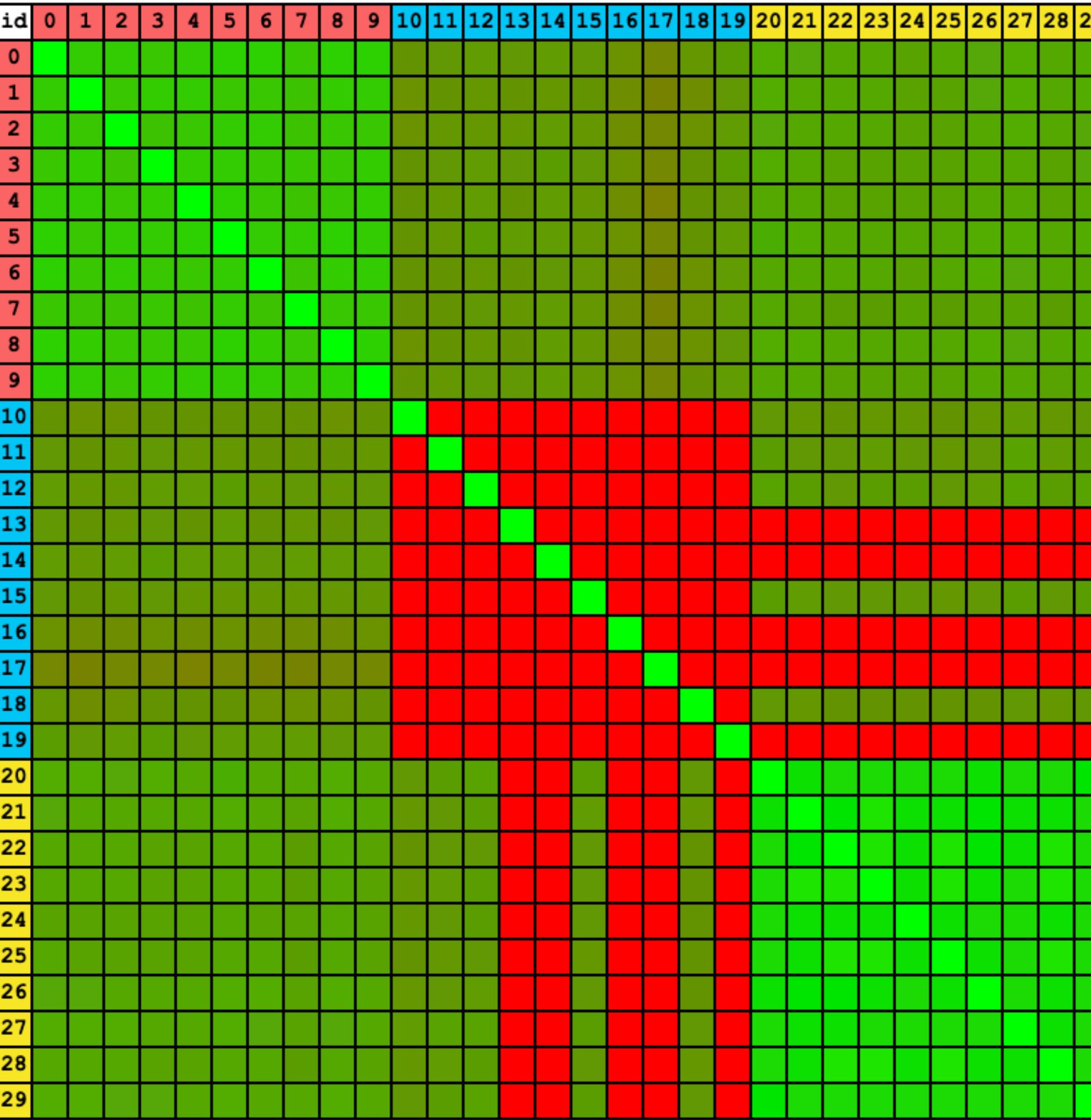
GDD agreement **geometric** comparison

group labels colors

BA *background:*  *text:* 

ER *background:*  *text:* 

GEO *background:*  *text:* 



Výsledky



$|V| = 100$

$|E| = 200$

umelo vytvorené siete

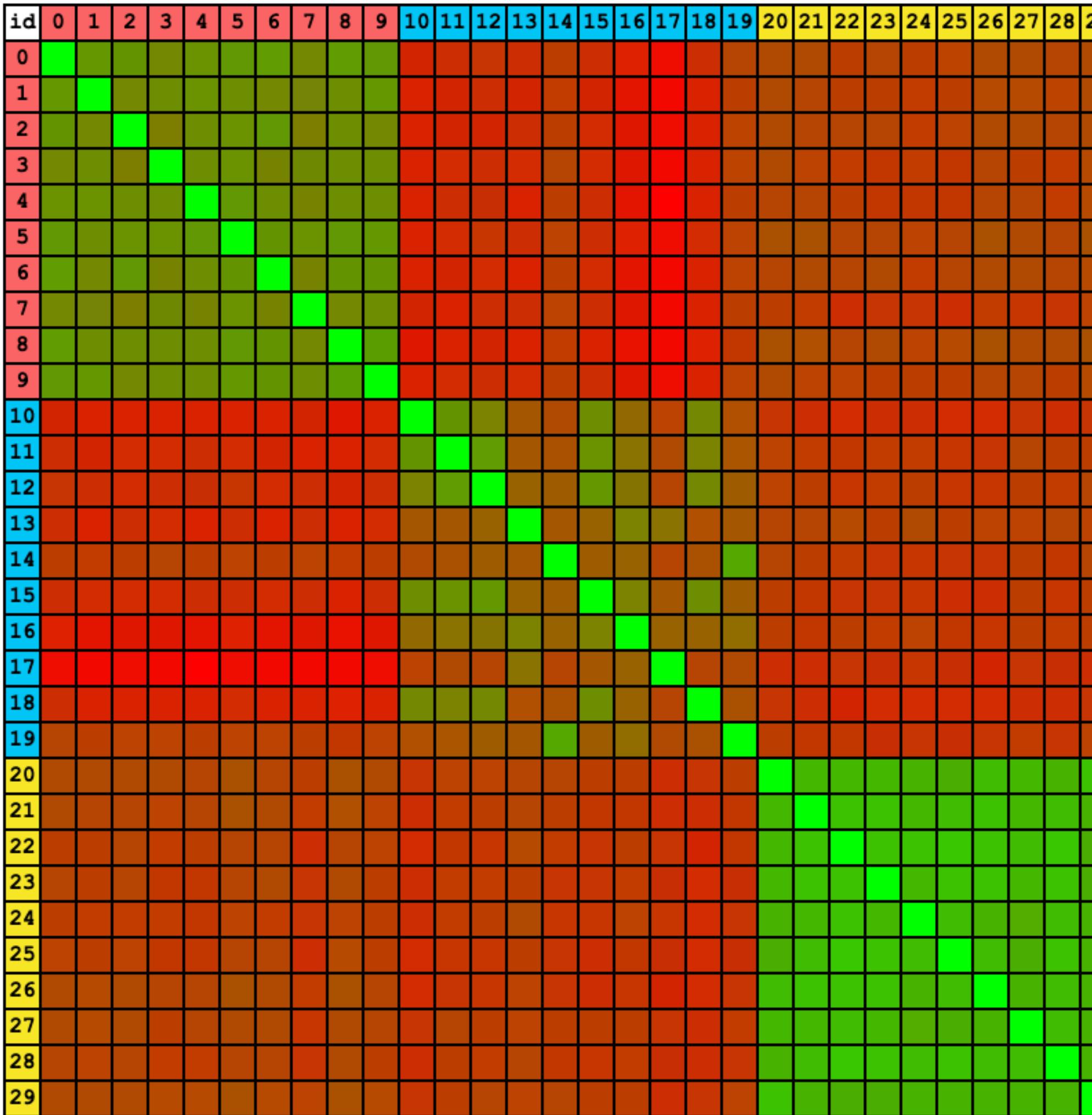
GDD agreement **arithmetic** comparison

group labels colors

BA *background:*  *text:* 

ER *background:*  *text:* 

GEO *background:*  *text:* 



Výsledky



$|V| = 100$

$|E| = 200$

umelo vytvorené siete

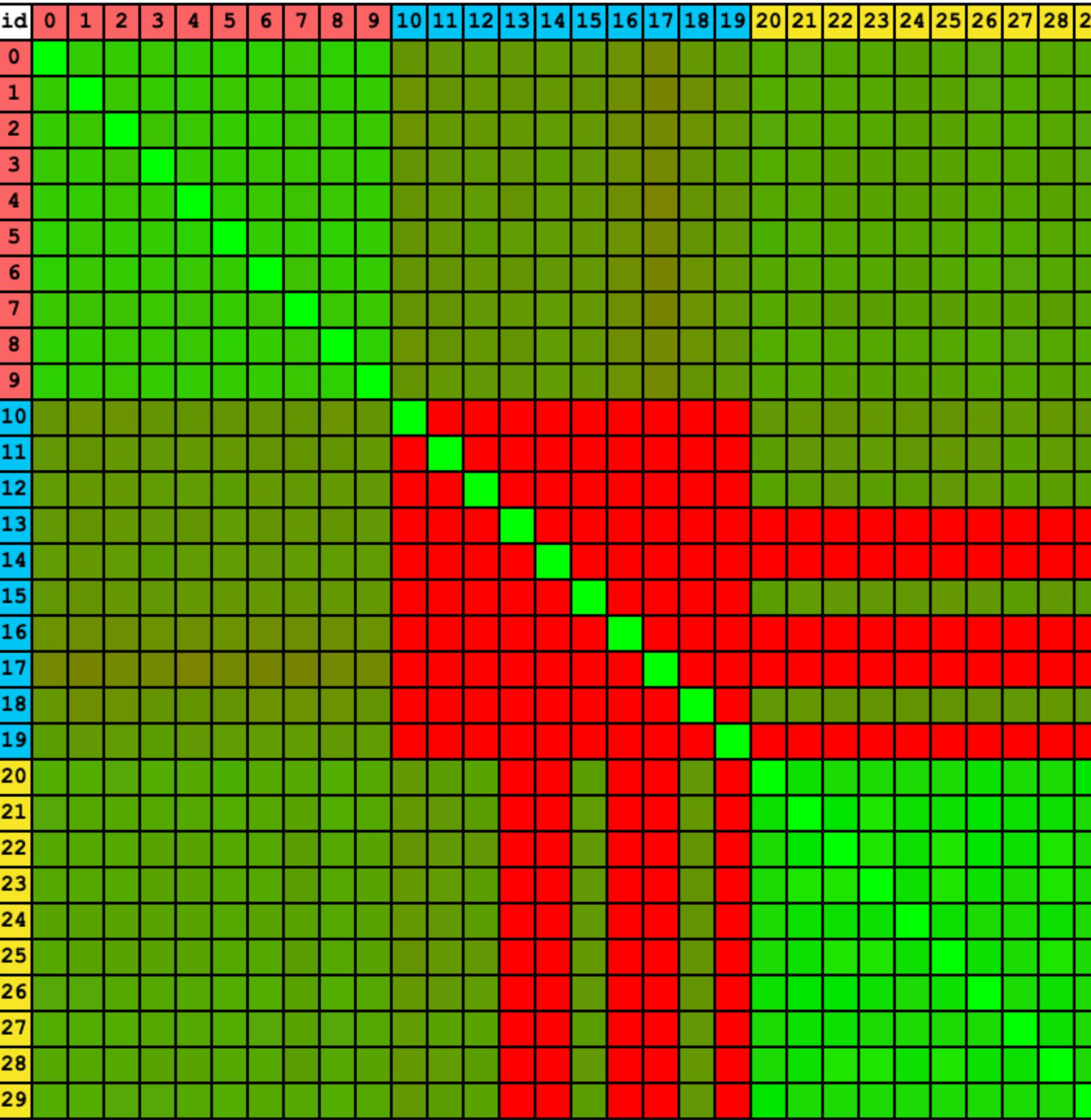
GDD agreement **geometric** comparison

group labels colors

BA *background:*  *text:* 

ER *background:*  *text:* 

GEO *background:*  *text:* 



Výsledky

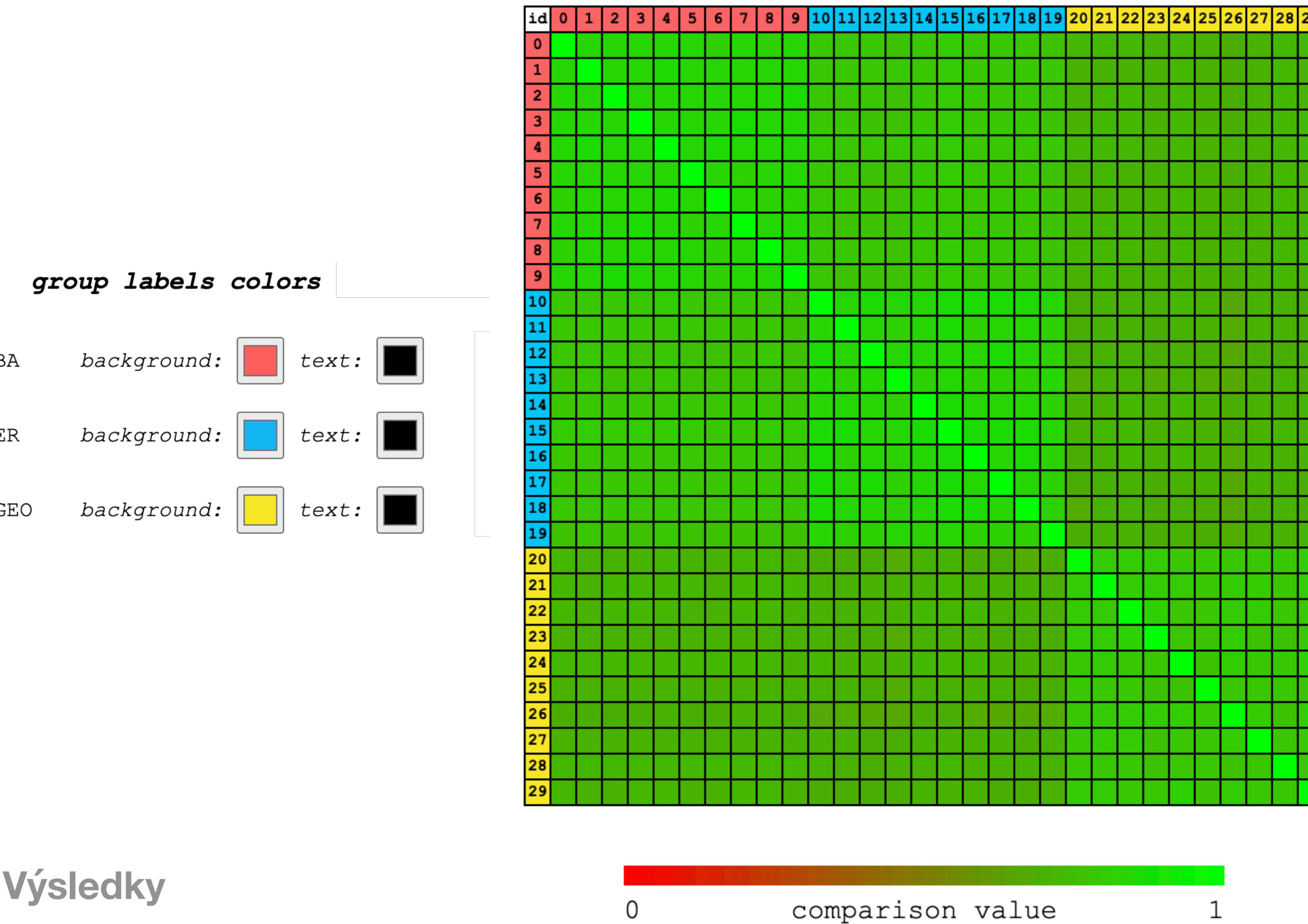


$|V| = 100$

$|E| = 200$

umelo vytvorené siete

GDD agreement **arithmetic** comparison



$|V| = 100$

$|E| = 500$

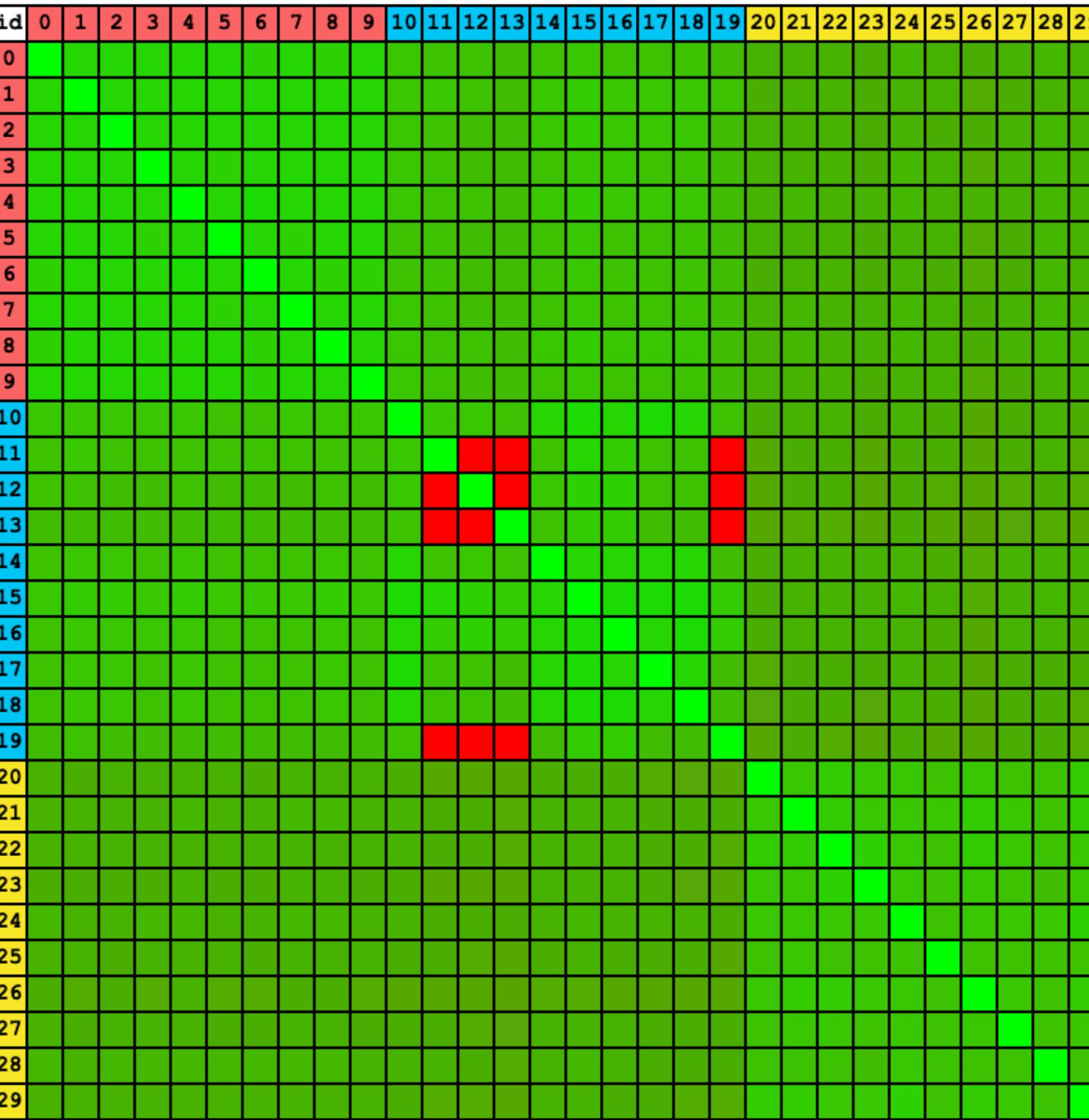
GDD agreement **geometric** comparison

group labels colors

BA *background:*  *text:* 

ER *background:*  *text:* 

GEO *background:*  *text:* 

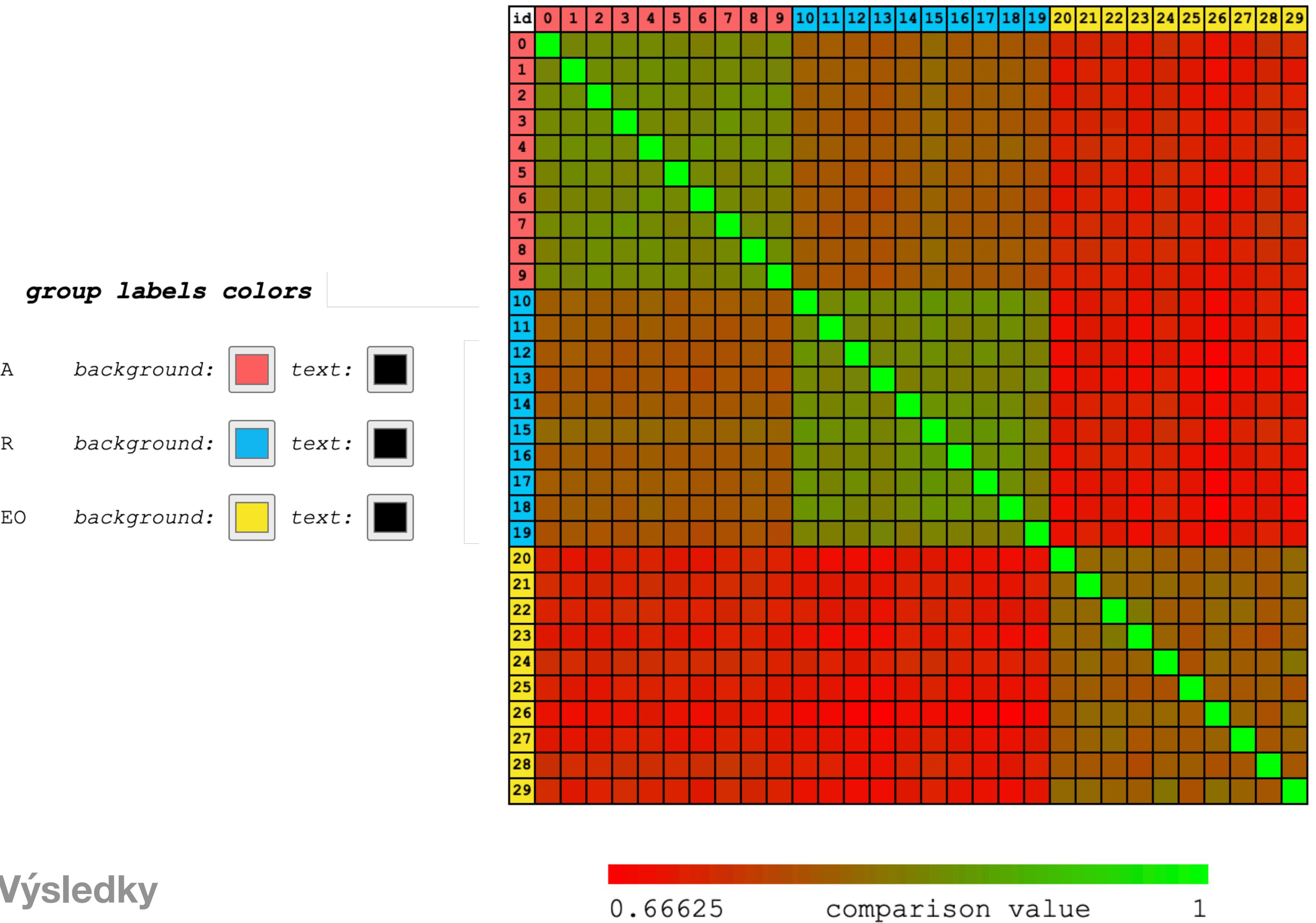


Výsledky

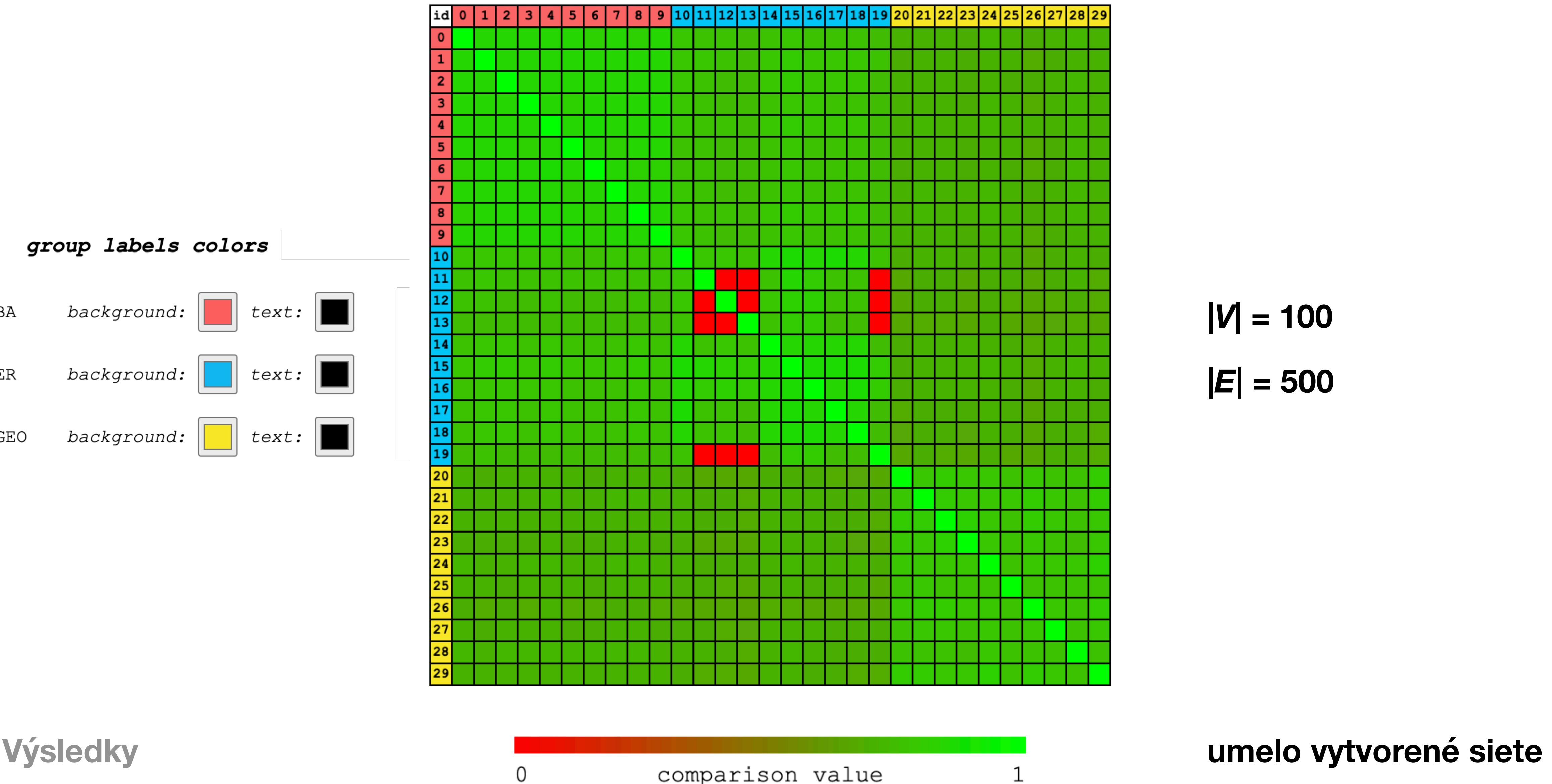


umelo vytvorené siete

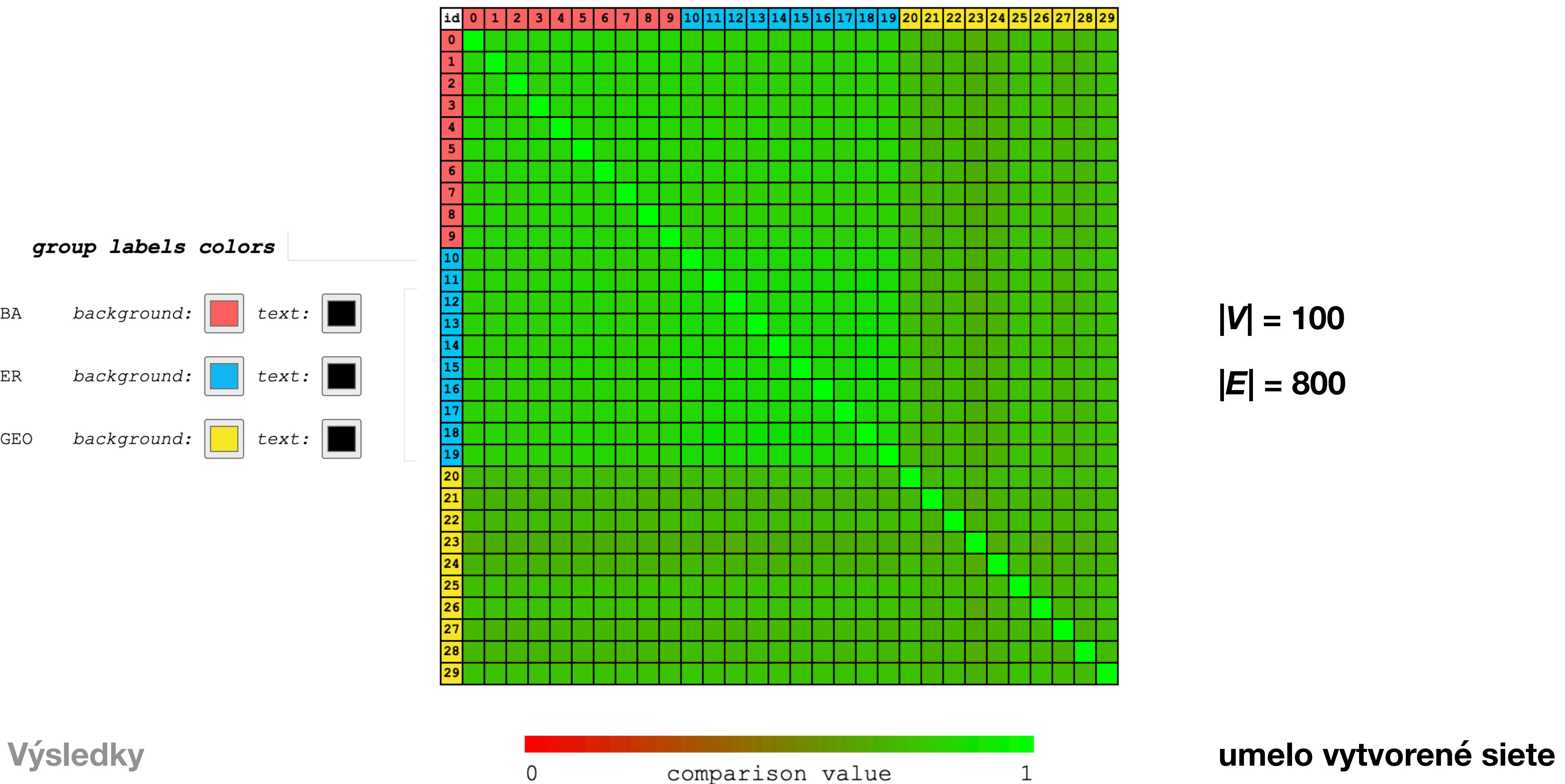
GDD agreement **arithmetic** comparison



GDD agreement **geometric** comparison



GDD agreement **arithmetic** comparison



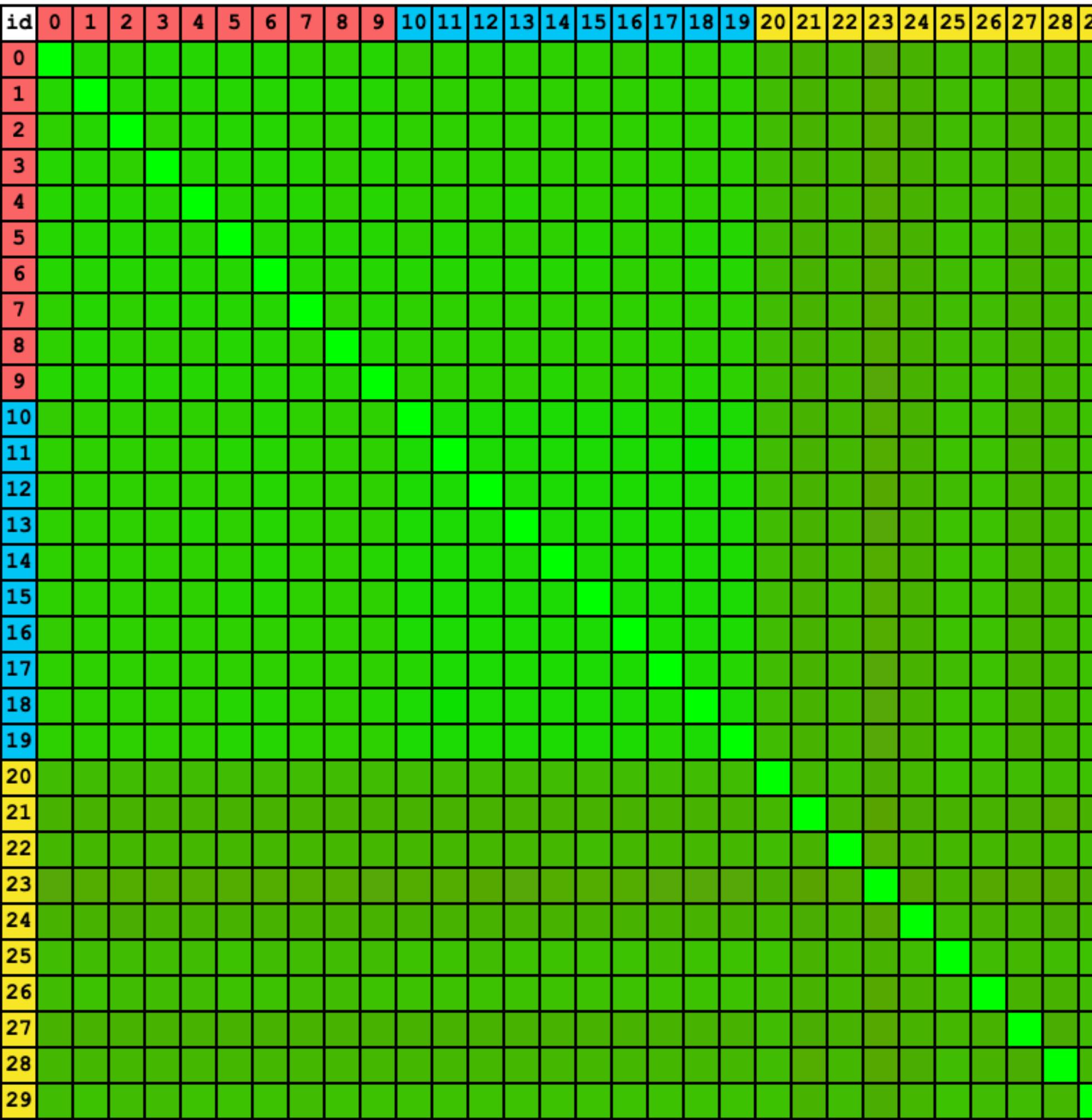
GDD agreement **geometric** comparison

group labels colors

BA *background:*  *text:* 

ER *background:*  *text:* 

GEO *background:*  *text:* 



Výsledky

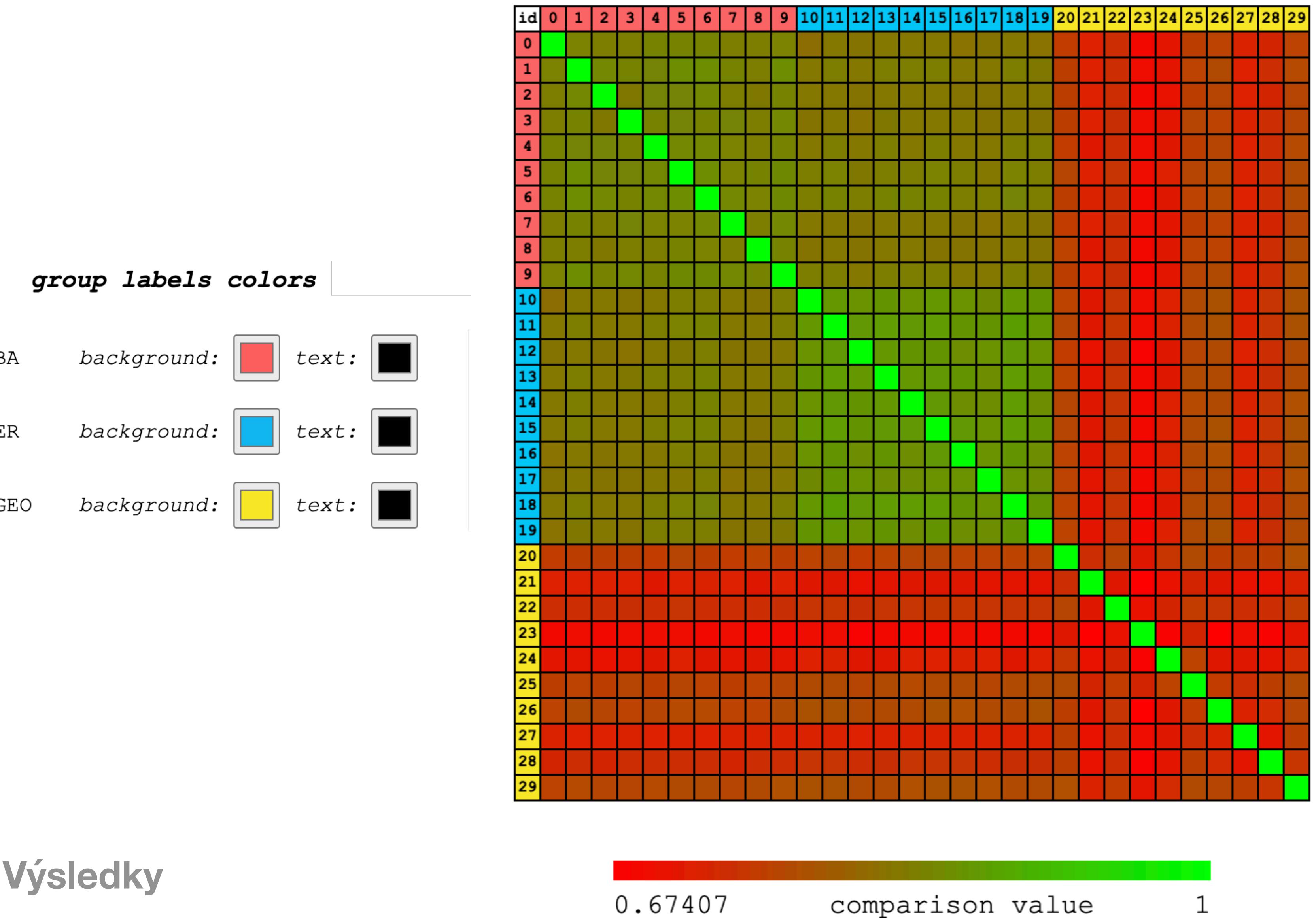


$|V| = 100$

$|E| = 800$

umelo vytvorené siete

GDD agreement **arithmetic** comparison



$|V| = 100$

$|E| = 800$

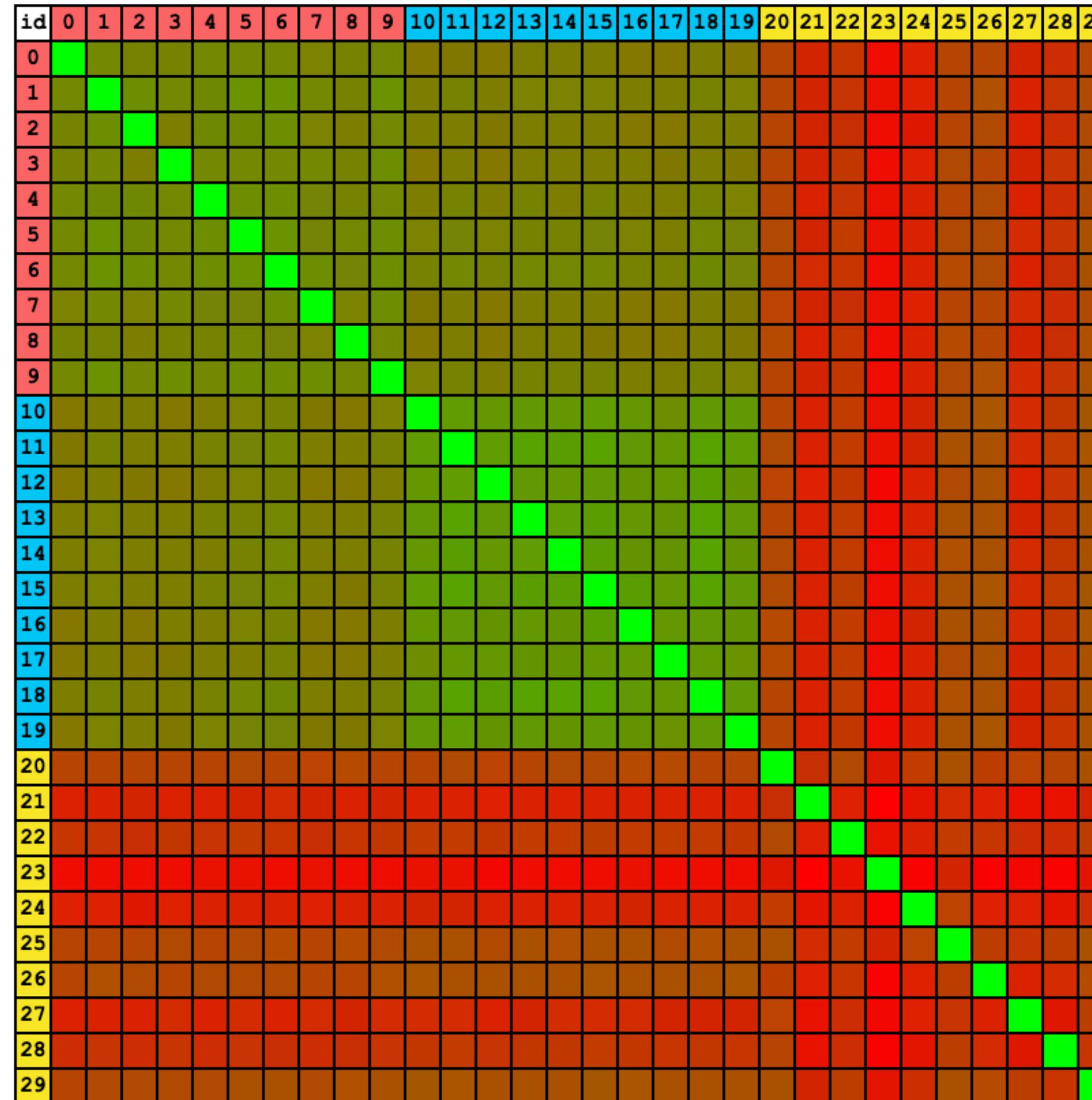
GDD agreement **geometric** comparison

group labels colors

BA *background:*  *text:* 

ER *background:*  *text:* 

GEO *background:*  *text:* 



Výsledky



$|V| = 100$

$|E| = 800$

umelo vytvorené siete

Zhrnutie

úspešne sme splnili všetky ciele

- Vytvorili sme software na počítanie GDD pomocou kombinatorického algoritmu ORCA.
- Vytvorili sme ľahko použiteľný, intuitívny software na počítanie a porovnávanie súhlasu GDD sietí. Náš software počíta súhlasy GDD s vysokou presnosťou.
- Úspešne sme overili súhlas GDD ako vhodnú mieru na porovnávanie štruktúry grafov rozlíšením skupín umelo generovaných BA, ER, GEO sietí.
- Kvôli zašumenosťi dát sa nám nepodarilo porovnávaním súhlasov GDD rozlíšiť skupinu pacientov s Alzheimerovov chorobov od zdravých jedincov.

Ďakujem za pozornosť

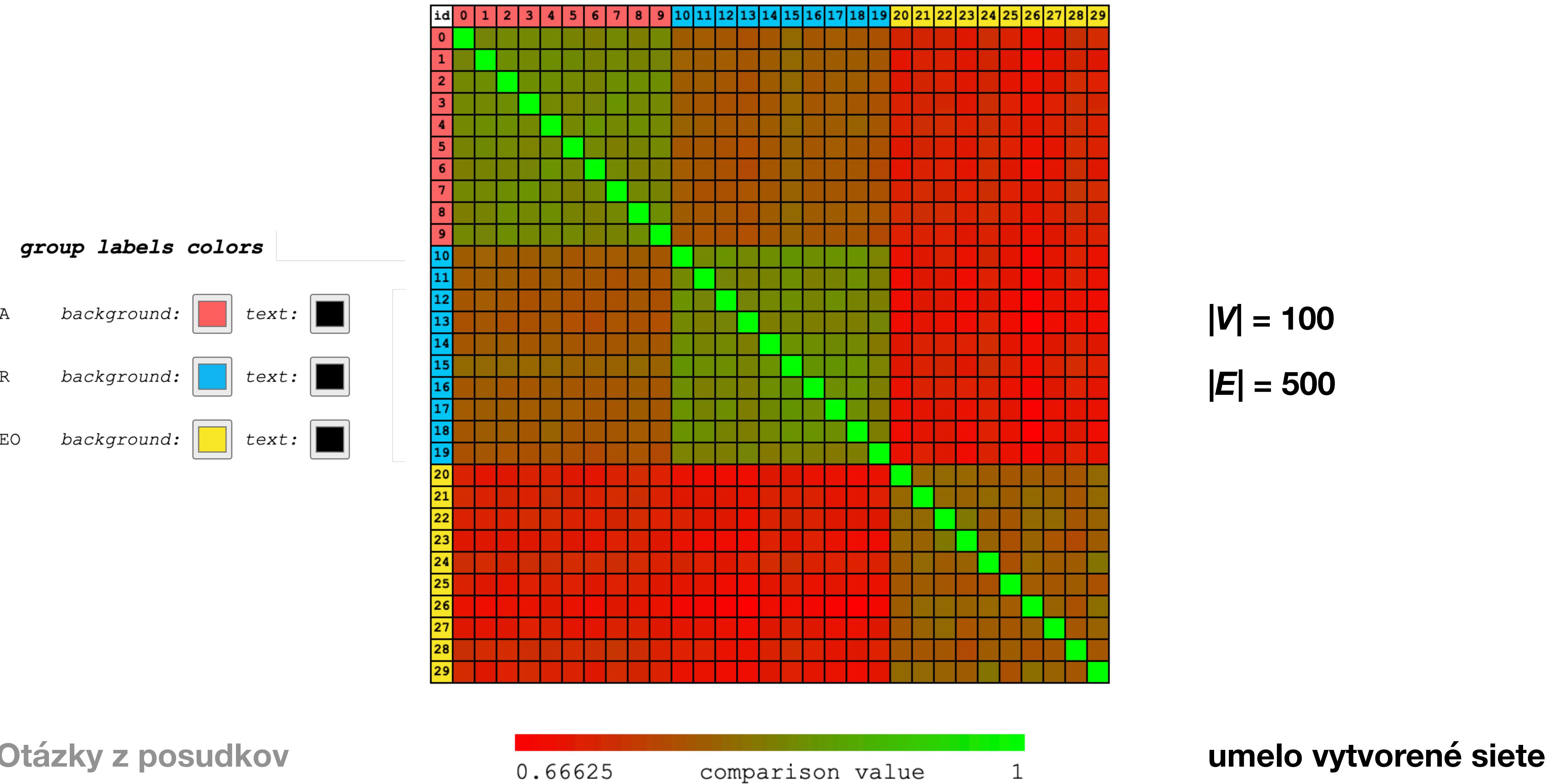
<https://github.com/michalpuskel/diplomka>

Otázky školiteľky

1. Ako by autor overil, či GDD súhlas dvoch grafov závisí od parametrov grafov?

Porovnal by som aj grafy s rovnakými parametrami, aby sa potvrdila / vyvrátila závislosť spôsobnosti porovnávania súhlasov GDD od parametrov grafov.

GDD agreement **arithmetic** comparison



Otázky školiteľky

2. Čo je možnou príčinou toho, že GDD súhlas nie je dobrou mierou pre porovnanie funkčných sietí mozgu zdravých osôb a osôb s Alzheimerom?

Šum v dátach. Ľudia dýchajú, hýbu sa trochu pri meraní a pod. Teda nezašumené dáta nie sú možné. Mozog tiež vytvára spontánne impulzy, napr. aj keď na nič nemyslíme, mozog pracuje. Čiže nezašumené dáta z princípu nie sú možné.

Ďalším problémom je rôzny počet vrcholov a hrán funkčných sietí mozgu. Siete sa nedajú znormalizovať na rovnaký počet vrcholov a hrán, lebo ľudské mozgy sú rôzne. Šum vzniká aj tým, že sa dáta premapovávajú na štandardizovanú veľkosť lebky, aby sa odstránil vplyv jej veľkosti.

Otázky školiteľky

3. Mohol by autor zdôrazniť, čo je najdôležitejším originálnym výsledkom diplomovej práce?

Vytvorený software na analýzu súhlasov GDD sietí. Software bol pôvodne zamýšľaný ako rozširujúci modul do už existujúcej aplikácie vytvorenej v bakalárskej práci. To bolo príčinou pokračovania vo výbere zvolenej technológie - programovacieho jazyka GO, aby sa zabezpečila spätná kompatibilita.

Neochvejná vášeň pre výskum, ale hlavne pre programovanie zapríčinila naše nezmierenie sa s neschopnosťou rozlísiť skupiny pacientov s Alzheimerovov chorobov, už dávno pred rokmi, keď všetko nasvedčovalo tomu, že pravdepodobne sa skupiny nedajú rozlísiť.

To viedlo následne k prepísaniu aplikácie na počítanie s číslami s arbitrárnou presnosťou. Žiaľ v tom čase v jazyku GO neexistovala implementácia matematickej operácie extrahovania n-tej odmocniny v knižnici pre počítanie s veľkými číslami.

My, ako správni programátori, sme sa problému nezľakli, ale sme algoritmus naštudovali, pochopili a naprogramovali tak, že funguje a výsledkom je unikátny software s vlastnou implementáciou počítania n-tej odmocniny, ktorý počíta s reálnymi číslami s presnosťou 1200 bitov v mantise (v defaultnom nastavení).

Otázky z posudkov

Otázky oponentky

1. V práci sa odvolávate na dizertačnú prácu doktora Andreja Jursu, ktorá, narozenie od vašej, popisuje úspešné využitie súhlasu GDD na rozlišovanie modelov komplexných sietí. Vysvetlite, v čom je rozdiel medzi jeho a vašimi experimentami a ako by bolo možné vaše experimenty upraviť tak, aby sa efekt potvrdil a prečo je tomu tak.

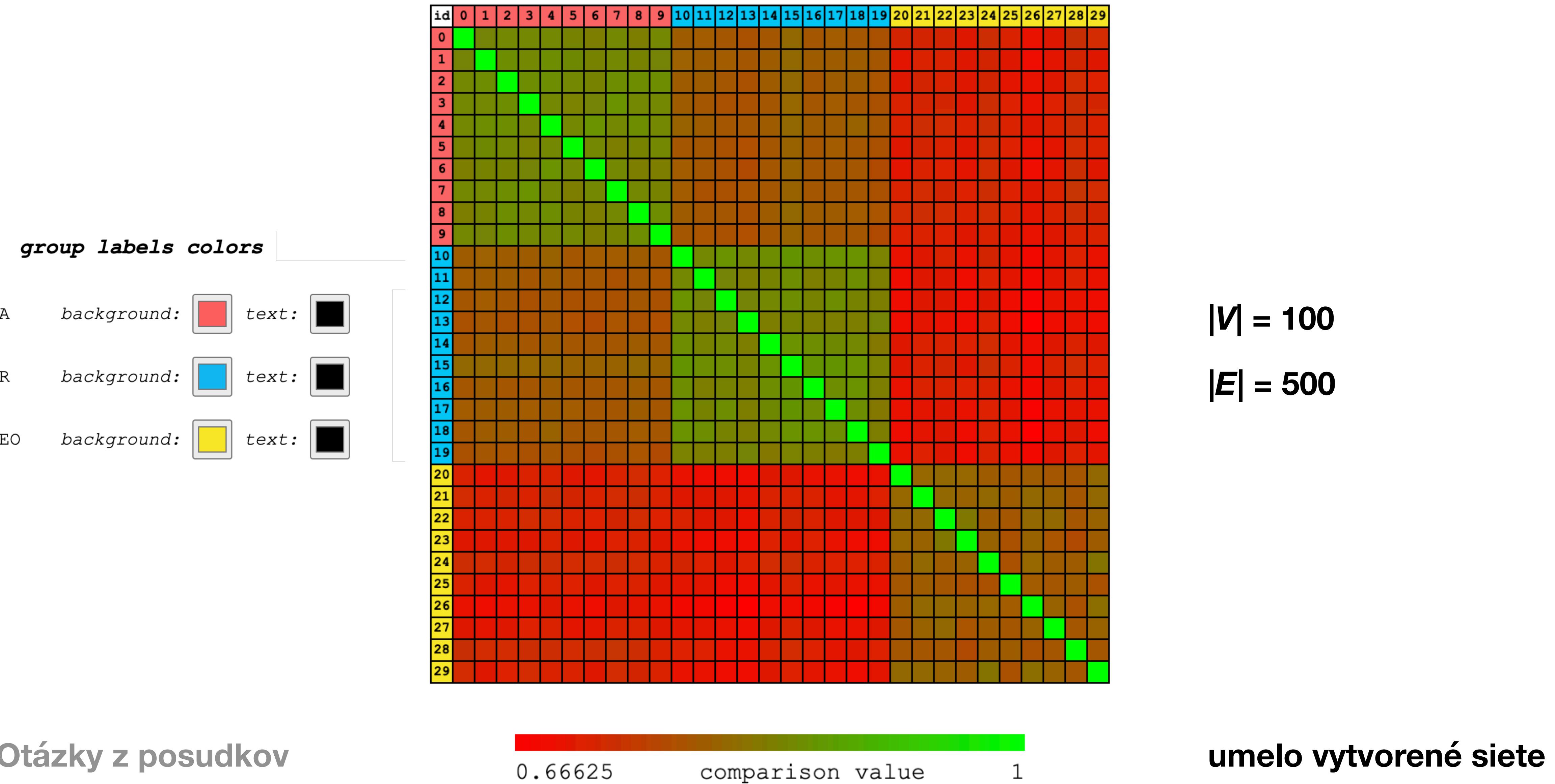
Doktor Andrej Jursa porovnával grafy s rovnakým počtom vrcholov a hrán. Toto je pre použitie GDD súhlasu klíčové. Naprogramovali sme si vlastné generátory na tvorbu BA, ER, GEO sietí tak, aby sme mohli porovnať 3 vzorky s približne rovnakým počtom vrcholov a hrán a výsledky preukázali, že GDD naozaj funguje.

Ked' sú grafy toho istého typu veľmi rozdielne, čo sa týka počtu uzlov a hrán, niektoré grafletry nemajú šancu vzniknúť. Ak má napr. nejaký graf 3 uzly, može tam vzniknúť graflet typu 0,1 a 2 ale nie ďalšie. Ak má 10000 uzlov, tak aj ďalšie.

Pôvodne sme sa domnievali, že normalizácia distribúcií vo vzorci GDD súhlasu ošetruje problém porovnávania grafov s rôznym počtom vrcholov a hrán, no v praxi to úplne nefunguje. Stále je ale GDD súhlas lepšia a rýchlejšia miera ako skúmať izomorfizmus grafov. Resp. normalizácia GDD asi funguje, ale pre velikánske grafy, kde sa už stráca vplyv ohraničenej veľkosti grafu.

Otázky z posudkov

GDD agreement **arithmetic** comparison



Otázky oponentky

2. V prípade funkčných sietí mozgu sa nenašiel žiadny signifikantný rozdiel súhlasu GDD medzi zdravými a chorými subjektami. Je možné, že to súvisí s kvalitou dát. Viete si predstaviť resp. navrhnuť ako možno predísť tomu, aby analýzu grafov ovplyvnila miera šumu či kvality dát? Je možné dáta spracovať tak, aby sa efekt našiel?

Najlepšie by bolo, keby boli dáta bez šumu lebo ten spôsobuje, že sa všetky grafy na seba podobajú lebo GDD súhlas v nich odhalí ten podobný šum. Keby sa dali identifikovať konkrétné grafletry resp. orbity v ktorých sa zašumené časti nenachádzajú, tak by sa dal upraviť GDD súhlas, aby porovnával iba orbity nášho záujmu.

Rozhodne však najlepšie by bolo získať nezašumené dáta. Bolo by treba ladiť korelačný prah aby BOLD vzorkovanie z fMRI dokázalo vygenerovať použiteľné vstupné grafy. Navyše dáta by mali byť znormalizované, aby mali približne rovnaký počet vrcholov a hrán.

Lenže keby sa to dalo, tak by sme mali bezšumové dáta, ale ono to nejde, lebo ľudia dýchajú, hýbu sa trochu pri meraní a pod. Teda nezašumené dáta nie sú možné. Mozog tiež vytvára spontánne impulzy, napr. aj keď na nič nemyslíme, mozog pracuje. Čiže nezašumené dáta z princípu nie sú možné.

Ani sa to nedá znormalizovať na rovnaký počet vrcholov a hrán, lebo ľudské mozgy sú rôzne. Šum vzniká aj tým, že sa dáta premapovávajú na štandardizovanú veľkosť lebky, aby sa odstránil vplyv jej veľkosti.

Otázky z posudkov

Otázky oponentky

3. V súvislosti s otázkou 2.: viete navrhnuť, stačí konceptuálne, aká iná miera podobnosti grafov by mohla ukázať rozdiely medzi zdravými a chorými jedincami?

Keby sme to vedeli, už by sme to navrhli. Štandardné metódy to rozlíšiť nevedia.

Skúmať miery napr. s počtom grafletov by mohlo byť zaujímavé, avšak oproti GDD súhlasu by sme pravdepodobne nedostali konkrétnejšie výsledky nakoľko je to všeobecnejšia miera než skúmanie konkrétnych orbít, ktoré sú súčasťou grafletov.

Možno by mohlo byť zaujímavé preskúmať napr. betweenness centrality prvých x vrcholov s najväčším stupňom vrchola. Betweenness centrality meria počet najkratších cest prechádzajúcich daným vrcholom. Pravdepodobne zdravý mozog by mal byť viac prepojený. Možno zdravý mozog by mal mať výrazne vyššie hodnoty betweenness centrality? Opäť by ale pravdepodobne šum skresľoval výsledky...

Otázky z posudkov

Ďakujem za pozornosť
Koniec

<https://github.com/michalpuskel/diplomka>