UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



PREDIKOVANIE NEPOZOROVANÝCH HRÁN V ZNAMIENKOVEJ VÁŽENEJ SIETI

DIPLOMOVÁ PRÁCA

2021 Bc. Alex Babiš

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

PREDIKOVANIE NEPOZOROVANÝCH HRÁN V ZNAMIENKOVEJ SIETI

DIPLOMOVÁ PRÁCA

Študijný program: Ekonomicko-finančná matematika a modelovanie
Študijný odbor: 1114 Aplikovaná matematika
Školiace pracovisko: Katedra aplikovanej matematiky a štatistiky
Vedúci práce: doc. RNDr. Beáta Stehlíková, PhD.

2021 Bc. Alex Babiš





Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Alex Babiš
Študijný program: ekonomicko-finančná matematika a modelovanie (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Predikovanie nepozorovaných hrán v znamienkovej váženej sieti
Prediction of unobserved edges in a signed weighted network

Anotácia: V niektorých prípadoch je realistické predpokladať, že sociálna sieť nie je známa úplne, t. j. niektoré hrany nepozorujeme. V literatúre boli navrhnuté rôzne algoritmy, ktoré predikujú tieto nepozorované hrany. Vybrané algoritmy sa použijú na konkrétne siete.

Vedúci: doc. RNDr. Mgr. Beáta Stehlíková, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Marek Fila, DrSc.

Dátum zadania: 17.12.2019

Dátum schválenia: 21.12.2019

prof. RNDr. Daniel Ševčovič, DrSc.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie: Chcel by som sa poďakovať svojej vedúcej práce doc. RNDr. Beáte Stehlíkovej, PhD. za jej čas, trpezlivosť, ochotu, cenné rady a pripomienky, ktorými mi uľahčila písanie tejto práce. Taktiež by som chcel poďakovať svojej rodine a priateľom, ktorí ma dokázali motivovať.

Abstrakt v štátnom jazyku

Babiš, Alex: Predikovanie nepozorovaných hrán v znamienkovej váženej sieti [Diplomová práca], Univerzita Komenského v Bratislave. Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: doc. RNDr. Beáta Stehlíková, PhD., Bratislava, 2021, 82:

Predikcia nepozorovaných alebo budúcich hrán v sieťach je jednou z aktuálnych tém, ktorou sa v uplynulých dvoch dekádach zaoberalo množstvo ľudí. Cieľom našej práce bude priblížiť využívane techniky, špeciálne metódy založené na lokálnej topológii sieti. Väčšina navrhnutých algoritmov sa zaoberá predikciou hrán v binárnych sieťach, v krajnom prípade vážených sieťach. My sa budeme preto snažiť práve techniky využívajúce topológiu siete zovšeobecniť, aby ich bolo možné a zároveň vhodné použiť pre siete znamienkovo váženého typu. Jedným spôsobom ako predikovať hrany v znamienkových sieťach by bolo jednoducho pripočítať dostatočne veľkú konštantu k váham hrán, dosiahnuť tým ich kladnosť a použiť existujúce prístupy. V práci sa venujeme korelačným sieťam, kde berieme kladné aj záporné hrany ako formu symetrickej informácie. Pričítanie konštanty a transformovanie korelačnej siete na váženú sieť by nám daný predpoklad, ako aj prácu so sieťou, mohol výrazne narušiť. Preto je lepšie zaviesť a pracovať s technikami, ktoré budú schopné s takouto formou informácie pracovať priamo. V záverečnej časti práce načrtneme algoritmus pre filtrovanie informácie z plnej siete, ktorý sa bude opierať o zovšeobecnené prístupy predikcie hrán a filter porovnáme s existujúcimi prístupmi, akými sú minimálna kostra a trojuholníkovo maximálne filtrovaný graf. Výsledky budeme porovnávať na základe podobností zhlukovania, ktoré jednotlivé filtrované siete vytvoria.

Kľúčové slová: predikcia hrán, znamienkové vážené siete, filtrované siete, TOPSIS, zhlukovanie

Abstract

Babiš, Alex: Prediction of unobserved edges in a signed weighted network [Master thesis], Comenius University in Bratislava. Faculty of mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: doc. RNDr. Beáta Stehlíková, PhD., Bratislava, 2021, 82

Prediction of unobserved and future edges in networks is one of the most recent topics that has been discussed frequently in the past two decades. The aim of our paper is to introduce used techniques, specially methods based on local topology of network. Most of proposed algorithms are dedicated to predict links in binary networks or in fewer cases weighted networks. For that matter, our aim is to generalize techniques based on local topology so they are more suitable and can be used for predicting links in signed weighted networks. One way to predict links in signed weighted networks is to add sufficiently big positive constant to weights of edges which transforms network so existing approaches are applicable. In our paper, we work with correlation networks where we take both positive and negative edges as some kind of symmetric information. Adding constant and transforming of correlation network to weighted network could disrupt given assumption as much as working with network itself. Better way to deal with correlation network is to introduce and work with techniques which will be able to deal with this kind of information directly. In final part of our work we will design algoritm for filtering information from complex network which will be based on generalized approaches for link prediction and we will compare it with existing filtering techniques such as minimum spanning tree or triangulated maximally filtered graph. The results will be compared based on similarities between clusterings that each filter forms.

Keywords: link prediction, signed weighted networks, filtered networks, TOPSIS, clustering

Obsah

Úvod ... 9

1 Dôležité pojmy z teórie sietí ... 11

1.1 Čo je to sieť ... 11

1.2 Rozdelenie sietí ... 11

1.2.1 Delenie podľa hodnôt váh ... 11

1.2.2 Delenie podľa orientácie hrán ... 13

1.2.3 Delenie podľa počtu množín vrcholov ... 14

1.2.4 Delenie podľa časovej závislosti siete ... 14

1.3 Finančné siete ... 15

1.4 Filtrovanie informácie zo sietí ... 17

1.4.1 Víťaz berie všetko ... 17

1.4.2 Minimálna kostra ... 18

1.4.3 Planárny graf ... 19

1.5 Koeficient zhlukovania ... 21

1.6 Centrality a centralizacia ... 22

2 Predikcia hrán ... 27

2.1 Prehľad rôznych prístupov ... 27

2.1.1 Metódy založené na podobnosti vrcholov ... 27

2.1.2 Pravdepodobnostné prístupy ... 28

2.1.3 Metódy strojového učenia ... 28

2.2 Lokálne metódy a ich zovšeobecnenie ... 29

2.2.1 Spoloční susedia ... 29

2.2.2 Jaccardov index ... 32

2.2.3 Saltonov index ... 33

2.2.4 Hub promoted index 34
2.2.5 Hub depressed index 35
2.2.6 Leicht-Holme-Newmanov index 35
2.2.7 Local affinity structure index 35
2.2.8 Adamic-Adarov index 36
2.2.9 Resource Allocation Index 38
2.3 Analýza fungovania zovšeobecnených indexov 39
2.3.1 Dáta 39
2.4 Porovnávanie indexov 42
3 Filtrovanie informácie zo siete 48
3.1 Algoritmus založený na podobnosti vrcholov 48
3.1.1 Algoritmus 49
3.2 Filter zohľadňujúci viacero indexov naraz 55
3.2.1 TOPSIS algoritmus 55
4 Porovnanie prístupov 62
4.1 Girvan-Newmanov algoritmus 62
4.2 Algoritmus Walktrap 63
4.3 Upravený Randov index 64
Záver 70
Zoznam použitej literatúry 73

Úvod

Predikcia hrán v komplexných sieťach je jednou z horúcich oblastí aplikovanej matematiky. Priťahuje pozornosť medziodborových vedeckých komunít, a to hlavne kvôli svojej rozsiahlej aplikácii. Niekoľko takýchto aplikácií z rôznych odvetví v krátkosti na začiatok uvedieme.

S rapídnym vývojom internetu a internetových služieb vznikajú platformy, ktoré ponúkajú široký sortiment služieb. Sociálne siete ako Facebook, Twitter, Instagram slúžia na uľahčenie komunikácie medzi ľuďmi na diaľku alebo na zdieľanie zážitkov medzi známymi. V týchto sieťach si môžeme ako vrcholy predstaviť konkrétnych ľudí a hrany ako interakcie medzi nimi. Predikcia hrán je dôležitá napríklad pri odporúčaní nových známych alebo pri odporúčaní, koho sledovať na základe už existujúcich spojení v sieti. Iný príklad služieb ponúkajú streamovacie platformy, akou je napríklad platforma Netflix. Tieto formujú špeciálny typ sietí, takzvané bipartitné, ktoré majú dve množiny vrcholov. Prvá množina vrcholov predstavuje používateľov a druhá predstavuje určitý objekt. V prípade Netflixu sú objektami filmy alebo seriály. Úloha predikcie hrán v takomto prípade spočíva vo vytvorení odporúčacieho systému, ktorý by používateľom odporučil určitý film alebo seriál na základe ich histórie pozerania. Takýmto aplikáciám sa venovali autori napríklad v [1], [2], [3]. Zaujímavú triedu sociálnych sietí, tvorí sieť spoluautorstva, ktorá je obľúbená v akademických kruhoch kvôli svojej jednoduchosti a teda je vhodná na testovanie nových prístupov v oblasti predikcie hrán.

Ďalšou vednou disciplínou, ktorá ťaží z predikcií hrán, je biológia. Tu sa stretávame sa zaujímavými sieťami, akými sú napríklad sieť interakcií proteínov, sieť génovej regulácie a génovej expresie, rôzne metabolické siete a podobne. Aplikácií predikcie hrán v týchto sieťach je viacero [4], [5], [6].

Špeciálnu triedu sietí tvoria korelačné siete. Ide o jednoduchú sieť, ktorej vrcholy predstavujú určité subjekty a hrany predstavujú korelácie medzi nimi. Korelačné siete sú typom znamienkových a vážených sietí, to znamená, že hrany môžu mať kladné aj záporné váhy. Sú taktiež plne prepojené, čo znamená, že každý vrchol v sieti je prepojený hranou so všetkými ostatnými vrcholmi. A práve toto spôsobuje, že dôležitá informácia zo siete nie je na prvý pohľad vidieť. Z tohto dôvodu vznikali rôzne filtre,

ktoré nám napomáhajú odokryť hlavnú štruktúru siete a nepotrebné hrany z nej odstrániť. Najrozšírenejším nástrojom na filtrovanie informácie zo siete je minimálna kostra.

Práve týmto sieťam sa v práci budeme bližšie venovať, a to najmä kvôli ich využitiu vo finančnom svete. Ich popularita v poslednej dobe rastie, a to hlavne z dôvodu, že predstavujú spôsob, ako získať zaujímavú informáciu o správaní, o štruktúre, o stabilite finančného sektora alebo o interakciách medzi „hráčmi" na akciových trhoch.

Našim cieľom bude zovšeobecniť metódy určené na predikciu hrán v nevážených sieťach, poprípade vo sieťach, ktoré majú striktne nezáporné váhy hrán, a s ich pomocou vytvoriť jednoduchý filter, ktorý by pomohol k odokrytiu širšej štruktúry siete.

Na záver úvodu uvedieme v krátkosti štruktúru práce. V prvej kapitole sa budeme venovať pre nás dôležitým pojmom z teórie sietí. Zadefinujeme si samotný pojem siete, prejdeme rozdelením sieti podľa množín vrcholov, typom hrán. Taktiež si uvedieme typy filtrovaných sietí a v poslednom rade si zadefinujeme pojmy ako koeficient zhlukovania a centrality. V druhej kapitole sa budeme venovať predikciám hrán v sieťach. Prejdeme si stručným prehľadom prístupov, ktoré v danej oblasti vznikli. Bližšie sa budeme venovať lokálnym indexom, ktoré následne zovšeobecníme aj pre znamienkové siete. V tretej kapitole skombinujeme metódy predikcie hrán spolu s filtrovaním informácie a vyskladáme jednoduchý algoritmus. Následne budeme analyzovať vývoj novovznikajúcich podsietí. Najskôr z pohľadu každého indexu samostatne. Neskôr využijeme algoritmus na viackriteriálne rozhodovanie TOPSIS na filtrovanie informácie vzhľadom na všetky indexy. V poslednej kapitole porovnáme všetky vzniknuté filtrované siete medzi sebou podľa analýzy zhlukov, ktoré v daných sieťach vzniknú. Na to budú použité dva algoritmy pre zhlukovanie v sieťach.

1 Dôležité pojmy z teórie sietí

1.1 Čo je to sieť

Analýza sietí nám pomáha získať hlbšie poznatky o štruktúre vzťahov v sociálnych sieťach, v štruktúre alebo procese zmeny v prírodných fenoménoch, dokonca aj v analýze biologických systémoch rôznych druhov organizmom, napríklad pri rozpoznávaní najvplyvnejších osôb v sieti a definovaní charakteristík skupín vrcholov. Čo teda možno chápať pod pojmom sieť?

Sieť predstavuje skupinu objektov (napríklad ľudí, krajiny, filmy ale aj proteíny, gény a podobne) a vzťahmi medzi nimi. V matematike sa niekedy môžeme stretnúť s pojmom graf namiesto siete, v tejto práci to bude predstavovať jedno a to isté. Štruktúra siete pozostáva z vrcholov a hrán, kde vrcholy reprezentujú objekty, ktoré chceme analyzovať a hrany reprezentujú vzťahy medzi nimi. Formálne budeme sieť zapisovať v tvare $G = \langle V, E \rangle$, kde množina $V = \{v_1, \dots, v_n\}$ je množina všetkých vrcholov a $E = \{e_1, \dots, e_m\}$ je množina hrán, pričom prvky tejto množiny označujú dvojice vrcholov, môžu byť aj usporiadané ak sa jedná o orientované siete, ak sú tieto vrcholy prepojené hranou [7]. Označením $|V|$ a $|E|$ bude rozumieť počet prvkov príslušných množín. Alternatívne môžeme definovať sieť pomocou matice susednosti, ktorú budeme označovať štvorcovú maticu $A = (a_{ij})$ rozmeru $|V| * |V|$. Takáto forma definovania siete bude veľmi užitočná vo zvyšku práce.

1.2 Rozdelenie sietí

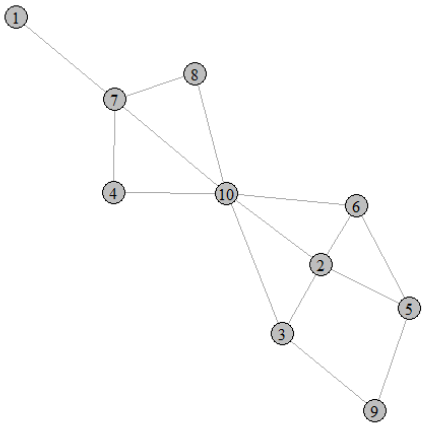
V závislosti na povahe vzťahov môžeme určiť rôzne kategórie delenia sietí.

1.2.1 Delenie podľa hodnôt váh

Prvú kategóriu delenia bude predstavovať delenia podľa hodnôt váh príslušných hrán v sieti.

Najjednoduchší prípadom je binárna sieť, obrázok 1. Táto sieť sa vyznačuje tým, že všetky hrany majú rovnakú váhu 1. Prvky matice susednosti sú

$$a_{ij} = \begin{cases} 1, & (v_i, v_j) \in E \\ 0, & (v_i, v_j) \notin E \end{cases}$$

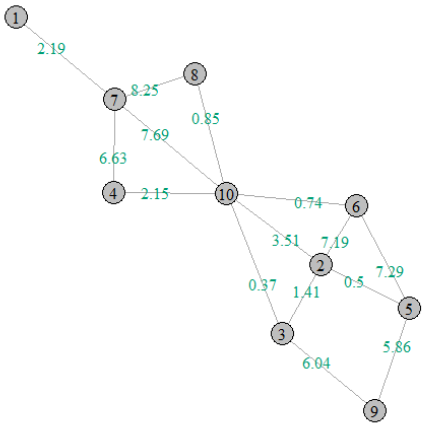


Obrázok 1 Binárna sieť

Druhým prípadom je vážená sieť, obrázok 2. V tomto prípade sú váhy hrán kladné reálne čísla. Prvky matice susednosti sú

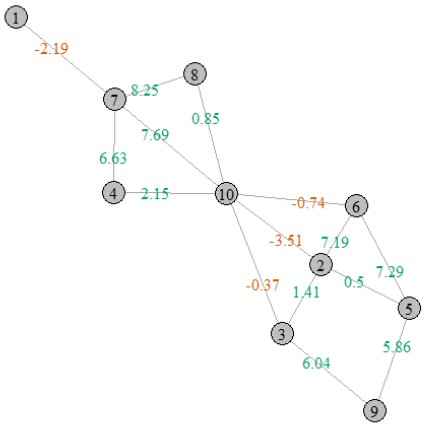
$$a_{ij} = \begin{cases} c_{ij}, & (v_i, v_j) \in E \\ 0, & (v_i, v_j) \notin E \end{cases}$$

$c_{ij} > 0$.



Obrázok 2 Vážená sieť

Ak by sme tento typ siete rozšírili aj o záporné váhy, t.j. $c_{ij} \epsilon \mathbb{R}$, dostali by sme znamienkové vážené siete, obrázok 3. Špeciálne ak sa obmedzíme iba na prípad $c_{ij} \in \{-1,1\}$, hovoríme iba o znamiekovej sieti. Vo všetkých prípadoch je matica susednosti symetrickou maticou.



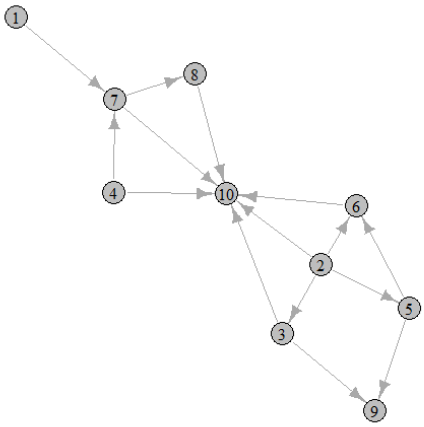
Obrázok 3 Znamienková vážená sieť

1.2.2 Delenie podľa orientácie hrán

Podľa orientácie hrán môžeme siete deliť na orientované alebo neorientované. Rozdiel medzi nimi je ten, že zatiaľ čo v prípade orientovanej siete chápeme význam hrany ako vzájomnú interakcie, v prípade neorientovanej siete to tak nemusí nutne byť. Orientácia hrany implikuje asymetrickosť vzťahov medzi dvoma aktérmi siete. Príkladom neorientovanej siete je napríklad sociálna sieť Facebook, kde hrany reprezentujú to, či sa dané dve osoby navzájom poznajú. Príkladom orientovanej siete je sociálna sieť Twitter, kde hrany reprezentujú to, kto koho sleduje, čo vo všeobecnosti nemusí byť vzájomné, obrázok 4. Matica susednosti pre orientované siete nie je symetrická a jej prvý sú

$$a_{ij} = \begin{cases} 1, & [v_i, v_j] \in E \\ 0, & [v_i, v_j] \notin E \end{cases}$$

kde $[v_i, v_j]$ označuje usporiadanú dvojicu vrcholov a chápeme pod tým, že hrana vedie z vrcholu v_i do vrcholu v_j.



Obrázok 4 Orientovaná sieť

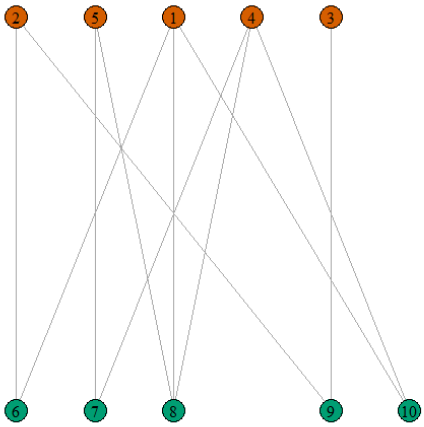
1.2.3 Delenie podľa počtu množín vrcholov

Podľa počtu množín vrcholov siete delíme na unipartitné, iba jedna množina vrcholov a multipartitné, viac množín vrcholov. Typickým príkladom multipartitných sietí sú takzvané bipartitné (dve množiny vrcholov) a tripartitné siete (tri množiny vrcholov). Pod bipartitnou sieťou si môžeme predstaviť streamovaciu platformu Netflix, kde jedna množina vrcholov predstavuje používateľov a druhá predstavuje filmy alebo seriály, obrázok 5. Hrany potom reprezentujú, či daný užívateľ videl alebo nevidel daný film. Pre bipartitné siete má matica susednosti rozmer $(|V_1| + |V_2|) * (|V_1| + |V_2|)$ a dá sa blokovo zapísať ako

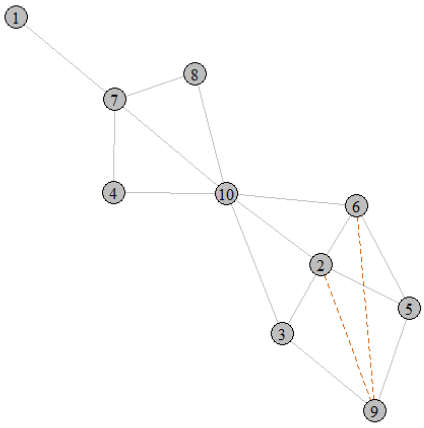
$$A = \begin{pmatrix} 0 & A_1 \\ A_2 & 0 \end{pmatrix}$$

1.2.4 Delenie podľa časovej závislosti siete

Posledné delenie sietí, ktoré uvedieme, je podľa ich časovej závislosti. Ak sa mení množina hrán alebo sa menia hodnoty váh hrán v závislosti od času, tak hovoríme o dynamických, obrázok 6, niekedy inak označovaných ako dočasných, sieťach.



Obrázok 5 Bipartitná sieť



Obrázok 6 Dynamická sieť - červenou sú znázornené budúce hrany

1.3 Finančné siete

Finančné siete ponúkajú zaujímavý prístup, ako nahliadnuť do štruktúr finančného sveta. Hlavnou myšlienkou je reprezentovať finančné subjekty ako vrcholy a ich empirickú závislosť ako váhy hrán medzi nimi, akou je napríklad korelácia.

Vrcholy môžu predstavovať napríklad krajiny, akcie alebo aj úrokové miery a hrany môžeme v jednotlivých prípadoch vyjadriť ako mieru korelácie medzi hrubým domácim produktom, výnosmi akcií, respektíve hodnotami úrokových mier. Prístupov, ako finančné siete skonštruovať, založených na rôznych typoch sietí, vzniklo viacero. Najčastejšie sa stretávame s korelačnými sieťami, ktoré predstavujú vážené neorientované siete. Väčšie hodnoty korelácie potom reprezentujú silnejšie spojenia vrcholov. Mantegna bol prvým, kto skonštruoval sieť z výnosov akcií použitím pearsonovho korelačného koeficientu [8]. Onnela a kol. [9] skonštruovali sieť akcií obchodovaných na Newyorskej burze rovnako pomocou korelačného koeficientu, a to tak, že začali s prázdnou sieťou a postupne pridávali hrany podľa poradia korelácií. V práci následne skúmali a porovnávali topologické vlastnosti vzniknutých podsietí. Tang a kol. [10] zvolili iný prístup. Rozhodli sa využiť namiesto korelačného koeficientu testovanie Grangerovej kauzality a zostrojili orientovanú sieť z 33 najväčších globálnych akciových indexov, pričom skúmali ako sa navzájom ovplyvňujú svetové trhy. Autori v [11] využili modely dynamickej korelácie, pričom skúmali vývoj trhu na základe dynamických korelačných sietí.

My sa v práci budeme zaoberať práve korelačnými sieťami, preto aj všetky príklady budú demonštrované na takýchto sieťach.

Niekedy budeme potrebovať pracovať namiesto váh hrán s vzdialenosťou vrcholov. Korelačný koeficient sa dá jednoducho previesť na euklidovskú vzdialenosť. Majme 2 štandardizované časové rady X a Y, čo znamená, že $E(X) = 0$, $E(Y) = 0$, $D(X) = 1$ a $E(Y) = 1$. Potom ich euklidovskú vzdialenosť môžeme vyjadriť nasledovne:

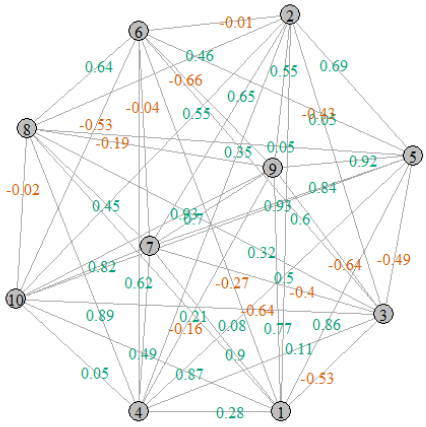
$$d^2(X,Y) = \sum_{i=1}^{n}(x_i - y_i)^2 = \sum_{i=1}^{n}(x_i^2 + y_i^2 - 2x_i y_i) = = \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2 - 2\rho(X,Y)$$
$$= 2n\big(1 - \rho(X,Y)\big)$$

Definujeme teda vzdialenosť medzi dvoma časovými radmi ako

$$d(X,Y) = \sqrt{2\big(1 - \rho(X,Y)\big)} \qquad (1)$$

1.4 Filtrovanie informácie zo sietí

Finančné siete predstavujú takzvané plné siete, obrázok 7. V praxi to znamená, že každá dvojica vrcholov je v sieti spojená hranou. Čím dostávame príliš vysoký počet hrán, ktorý je nutné zredukovať kvôli uľahčeniu skúmania charakteristiky sietí tak, aby sa zachovala základná informácia zo siete. Existujú tri najpoužívanejšie prístupy, víťaz berie všetko, minimálna kostra a planárny graf.



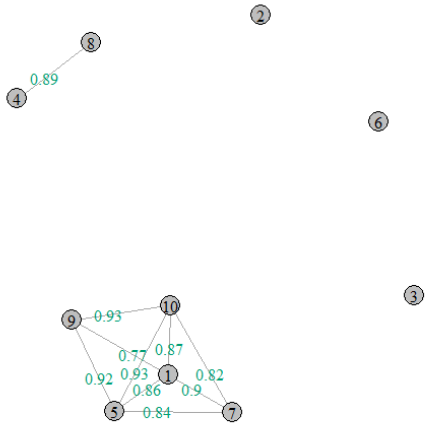
Obrázok 7 Korelačná sieť

1.4.1 Víťaz berie všetko

Ako aj samotný názov napovedá, ide o metódy pri ktorej sa určí hranica $t \in \langle -1,1 \rangle$, a všetky hrany s menšou váhou ako daná hranica budú zo siete eliminované. Aplikovaním dostávame filtrovanú maticu susednosti A s prvkami A_{ij} v tvare

$$A_{ij,(i\neq j)} = \begin{cases} 0, & \rho(x_i, x_j) < t \\ \rho(x_i, x_j), & \rho(x_i, x_j) \geq t \end{cases}$$

Takýto prístup využili autori v súvislosti s korelačnými sieťami v [12], [13], [14], [15]. Ako nevýhodu tejto metódy vidíme to pri malej hodnote hranice t nám v redukovanej sietí ostane stále veľa hrán a pri vyššej hodnote t sa nám sieť rozpadne na podsiete, obrázok 8.



Obrázok 8 Filtrovaná sieť z korelačnej matice -
metóda víťaz berie všetko s hranicou 0.7

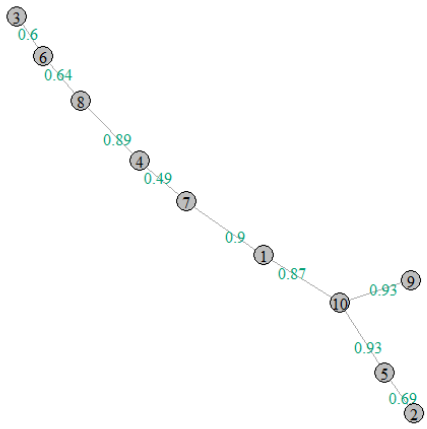
1.4.2 Minimálna kostra

Ďalším veľmi využívaným spôsobom na redukovanie počtu hrán v sieti je zostavenie minimálnej kostry (MST). Vzniklo množstvo efektívnych algoritmov ako minimálnu kostru grafu zostrojiť. Odporúčame čitateľovi prácu [16], ktorá preskúmava a porovnáva práce, ktoré sa zaoberajú problémom minimálnej kostry, motiváciu za ich vznikom a dáva ich do súvislosti s pokrokmi, ktoré sa v tejto oblasti spravili. V kontexte s korelačnými sieťami je práve minimálna kostra veľmi obľúbený nástroj, ako dolovať dôležitú informáciu zo siete [17], [18].

Hlavnou myšlienkou je vytvoriť súvislú sieť tak, aby celkový súčet váh všetkých hrán bol čo najmenší a zároveň daná sieť obsahovala všetky vrcholy a neobsahovala žiaden cyklus. Ak má sieť n vrcholov, potom po aplikovaní dostávame súvislú sieť s počtom hrán $n-1$. Minimálna kostra grafu je elegantný spôsob ako dostať zredukovanú sieť. Taktiež odstraňuje nedostatok metódy víťaz berie všetko tým, že výsledkom je súvislá sieť. Avšak stále vidíme potencionálne nedostatky v podobe počtu hrán, ktorý je nižší ako by sme si v ideálnom prípade možno želali.

Algoritmus minimálnej kostry je veľmi jednoduchý. Označme si $S_{ord} = (s_1, s_2, s_3, \dots)$ zoradenú postupnosť všetkých hrán vzostupne podľa vzdialeností jednotlivých dvojíc vrcholov. Začneme tým, že vyberieme hranu medzi dvojicou vrcholov, ktorých vzdialenosť je minimálna, teda s_1. Postupne budeme prechádzať zvyšné členy S_{ord} a hranu s_i pridáme iba vtedy ak dosiahnutý graf je stromom, neobsahuje cyklus, obrázok 9.

Ak chceme vytvoriť minimálnu kostru z korelačnej siete, nemožno teda priamo pracovať s váhami hrán. Preto v prípade minimálnej kostry pracujeme s maticou vzdialeností definovanou vzťahom (1).



Obrázok 9 Filtrovaná sieť z korelačnej matice -
metóda minimálnej kostry grafu

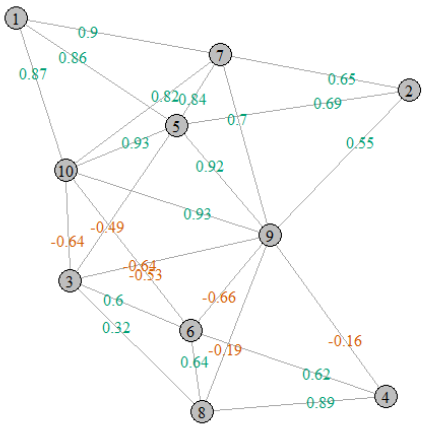
1.4.3 Planárny graf

Vylepšením minimálnej kostry je takzvaný planárne maximálny filtrovaný graf (PMFG) [19]. Výsledkom algoritmu je takzvaný planárny graf, čo znamená, že sa dá zakresliť v rovine tak, že hrany majú prienik iba na ich koncoch. V súvislosti s korelačnými sieťami uvádzame dielo [20], kde sa autori venovali skúmaniu topologických vlastnostiam filtrovanej siete, ktorá vznikla z 300 najobchodovanejších akcií na Newyorskej burze.

Algoritmus konštrukcie PMFG bude veľmi podobný ako na konštrukciu minimálnej kostry. Opäť si označme $S_{ord} = (s_1, s_2, s_3, \dots)$ zoradenú postupnosť všetkých hrán vzostupne podľa vzdialeností jednotlivých dvojíc vrcholov. Rovnako začneme tým, že vyberieme hranu medzi dvojicou vrcholov, ktorých vzdialenosť je minimálna, teda s_1. Taktiež budeme postupne prechádzať všetky členy S_{ord} s tým rozdielom, že tento krát pridáme hranu s_i iba vtedy ak je výsledný graf planárny. Je všeobecné potvrdené [21], že daný graf má počet hrán rovný $3n - 6$. Taktiež minimálna kostra je vždy súčasťou PMFG.

Vylepšenie oproti minimálnej kostre je pridanie informácií zvýšením počtu hrán. Naopak nevýhodou je pomalosť daného algoritmu, a taktiež nie je isté, prečo by práve planarita mala byť dobrým predpokladom pre filtrovanú sieť. Práve kvôli pomalému algoritmu sa častejšie používa približné riešenie v podobe trojuholníkovo maximálne filtrovaného grafu (TMFG) [22], obrázok 10.

Poslednú časť tejto kapitoly bude tvoriť priblíženie si dvoch pojmov z analýzy sociálnych sietí, ktoré budú veľmi nápomocné v neskoršej časti práce.



Obrázok 10 Filtrovaná sieť z korelačnej matice - metóda TMFG

1.5 Koeficient zhlukovania

Pre neorientovanú a neváženú sieť koeficient zhlukovania (CC) kvantifikuje početnosť spojených trojuholníkov v sieti [23]. Dá sa považovať za mieru, ktorá nám vyjadruje ako veľmi sú poprepájaní susedia ľubovoľného vrcholu medzi sebou. Inými slovami, koeficient zhlukovania sa počíta ako pomer počtu všetkých trojuholníkov, ktorých súčasťou daný vrchol a počtu všetkých trojíc centrovaných okolo daného vrcholu.

$$CC(v_x) = \frac{\sum_{i,j,i\neq j} a_{ix} * a_{jx} * a_{ij}}{k_{v_x} * \left(k_{v_x} - 1\right)}$$

kde a sú prvky matice susednosti a k_{v_x}je počet susedov vrcholu v_x. Z definovania koeficientu zhlukovania je zrejmé, že $CC(v_x) \in \langle 0,1 \rangle$. Hodnota $CC(v_x) = 0$ nám hovorí, že ani jeden zo susedov vrcholu v_x nie je prepojený hranou s iným susedom vrcholu v_x. Naopak $CC(v_x) = 1$ nám hovorí, že všetci susedia vrcholu v_x sú medzi sebou poprepájaní hranou. Globálny koeficient zhlukovania celej siete je potom iba priemerom hodnôt koeficientov zhlukovania všetkých vrcholov danej siete.

$$CC = \frac{1}{|V|}\sum_{i=1}^{|V|} CC(v_i)$$

Rovnako platí, že $CC \in \langle 0,1 \rangle$.

V našom prípade je to mierne komplikovanejšie. Keďže budeme ďalej pracovať s korelačnými sieťami budeme potrebovať zovšeobecnený koeficient zhlukovania pre znamienkové vážené siete, ktorý bude brať do úvahy aj váhy hrán a rovnako aj ich znamienko. Vzniklo niekoľko prístupov, ako toto dosiahnuť [24], ktoré boli prevažne odvodené pre koeficienty zhlukovania pre vážené siete. Prvým je koeficient zhlukovania daný nasledovným predpisom

$$CC_1(v_x) = \frac{\sum_{i,j,i\neq j} \left(w_{ix} * w_{jx} * w_{ij}\right)^{1/3}}{k_{v_x} * \left(k_{v_x} - 1\right)}$$

kde w_{ij} predstavuje váhu hrany medzi vrcholmi v_j a v_j. Ľahko sa dá vydedukovať, že tento krát je $CC_1(v_x) \in \langle -1,1 \rangle$. Okrajová hodnota $CC(v_x) = 1$ nám hovorí, že všetky dvojice susedov vrcholu v_x formujú s vrcholom v_x kladné trojuholníky (párny počet

kladný hrán v trojuholníku) a zároveň váhy všetkých prepojení dvojíc vrcholov v susedstve v_x sú v absolútnej hodnote 1. Opačný prípad $CC(v_x) = -1$ nám zase hovorí, že všetky dvojice susedov vrcholu v_x formujú s vrcholom v_x záporné trojuholníky (nepárny počet záporných hrán v trojuholníku) a zároveň váhy všetkých prepojení dvojíc vrcholov v susedstve v_x sú v absolútnej hodnote 1. Hodnotu $CC(v_x) = 0$ nadobúda v prípade, že všetky záporné a kladné trojuholníky v susedstve v_x sú vyvážené alebo žiaden zo susedov vrcholu v_x nie je spojený hranou s iným susedom vrcholu v_x.

Druhý zovšeobecnený koeficient zhlukovania je daný predpisom

$$CC_2(v_x) = \frac{\sum_{i,j,i\neq j} w_{ix} * w_{jx} * w_{ij}}{\sum_{i,j,i\neq j} |w_{ix} * w_{jx}|}$$

Rozdiel oproti $CC_1(v_x)$ je taký, že zatiaľ čo v pri indexe $CC_1(v_x)$ sa bral do úvahy iba počet možných trojuholníkov v susedstve v_x, v prípade $CC_2(v_x)$ sa berie do úvahy najväčšia možná intenzita daného trojuholníka. Rovnako ľahko ako v prípade $CC_1(v_x)$ sa dá uvidieť, že aj $CC_2(v_x)$ je z intervalu $\langle -1,1 \rangle$. Rovnako ako v prípade nevážených sietí definujeme globálny koeficient zhlukovania ako priemer koeficientov zhlukovania pre všetky vrcholy.

$$CC_1 = \frac{1}{|V|} \sum_{i=1}^{|V|} CC_1(v_i)$$

$$CC_2 = \frac{1}{|V|} \sum_{i=1}^{|V|} CC_2(v_i)$$

V oboch prípadoch platí, že $CC_i \in \langle -1,1 \rangle$. V práci budeme ďalej používať iba $CC_2(v_i)$ a CC_2.

1.6 Centrality a centralizacia

Centrality tvoria pomerne dôležitú časť nástrojov pre lepšie porozumenie siete. Kvôli mierne heterogénnej štruktúre sietí môžeme niektoré vrcholy považovať za dôležitejšie ako ostatné. Napríklad v sociálnych sieťach celebrity alebo politici majú veľa sledovateľov a môžu ľahšie propagovať, respektíve šíriť určité informácie oproti

bežným ľudom. Z finančného sveta zasa môžu niektoré obrovské firmy svojou aktivitou ovplyvňovať pohyb finančných trhov výraznejšie ako menšie firmy. Avšak definícia centrality nie je jednoznačná. Môžeme ju definovať z hľadiska „záťaže“, ktorú prijíma každý vrchol. Je silne závislá od aplikácie. Neexistuje konsenzus ako centralitu definovať, preto bolo navrhnutých niekoľko mier centralít, kde každá berie v úvahu špecifické aspekty. Pri definovaní všetkých spomenutých centralít, s výnimkou harmonickej centrality, budeme postupovať podľa [25] a v prípade harmonickej centrality podľa [26].

Najjednoduchšou centralitou je stupeň vrchola. Veľmi ľahko sa dá definovať ako počet vychádzajúcich hrán z vrcholu. Stupeň vrcholu sa nepozerá na váhy hrán. Je preto rovnaká pre vážené ako aj nevážené siete. Matematicky, pre vrchol v_x, ju budeme zapisovať ako

$$deg(v_x) = \sum_{v_i \in \Gamma(v_x)} 1$$

kde symbolom $\Gamma(v_x)$ označujeme množinu všetkých vrcholov, ktoré sú s vrcholom v_x spojené hranou.

Ak by sme chceli brať do úvahy váhy hrán, je lepšie použiť silu vrcholu namiesto jeho stupňa. Tá sa jednoducho definuje ako súčet váh všetkých hrán, ktoré vychádzajú z vrcholu. Pre vrchol v_x môžeme definovať silu ako

$$str(v_x) = \sum_{v_i \in \Gamma(v_x)} w_{ix}$$

Centralita vrcholov môže byť definovaná aj vzhľadom na vzdialenosti vrchola a nie váhy hrán. Potom vieme centrálny vrchol chápať ako vrchol, ktorý je blízko ku všetkým ostatným vrcholom v sieti. Táto myšlienka je zapracovaná v centralite blízkosti. Tá sa definuje pre vrchol v_x ako súčet najkratších ciest z vrcholu v_x do všetkých ostatných vrcholov v sieti. Interpretovať sa dá ako priemerná vzdialenosť vrcholu od všetkých ostatných. Centralitu blízkosti budeme zapisovať ako

$$clo(v_x) = \frac{|V|}{\sum_{i=1, i \neq x}^{|V|} d(v_x, v_i)}$$

pričom $d(v_x, v_i)$ bude označovať najkratšiu vzdialenosť medzi vrcholom v_x a v_i. Výhodou centrality blízkosti, ako aj stupňa a sily vrcholu, je, že je veľmi intuitívna a vhodná, ak chceme charakterizovať prenos informácie pozdĺž najkratšej vzdialenosti. Nevýhodou je, že ak je sieť nesúvislá, potom vzdialenosť medzi dvoma vrcholmi z rôznych podsietí je rovná nekonečnu, to znamená, že minimálne jedna najkratšia cesta pre každý vrchol v sieti bude nekonečno. Tým pádom je centralita blízkosti pre každý vrchol nulová a prakticky pre takúto sieť nepoužiteľná.

Tento problém rieši harmonická centralita, niekedy označovaná aj harmonická blízkosť. Harmonický priemer dokáže pracovať veľmi pohodlne aj s nekonečnými hodnotami, a preto sa stal vhodným kandidátom na riešenie problému ak sa klasická priemerná hodnota, ako v prípade centrality blízkosti, nedá použiť. Nahradíme teda priemernú vzdialenosť harmonickým priemerom, čo vedie k definície harmonickej centrality v tvare

$$clo(v_x) = \frac{1}{|V| - 1} \sum_{i=1, i \neq x}^{|V|} \frac{1}{d(v_x, v_i)}$$

Ak by sme sa chceli pozrieť na centralitu vrcholu vzhľadom na jeho vyťaženie v sieti, vhodného kandidáta by predstavovala centralita medzipolohy. Rovnako ako pri centralite blízkosti môžeme uvažovať, že nás zaujíma prenos informácie pozdĺž najkratšej cesty. Lenže vrcholy v sieti môžu mať aj viac ako len jednu najkratšiu cestu. V takom prípade je dobré sa na centrálne vrcholy pozerať z pohľadu ich zastúpenia v najkratších cestách medzi ostatnými vrcholmi. Matematicky centralitu medzipolohy definujeme ako

$$bet(v_x) = \sum_{i \neq j \neq x \in V} \frac{\eta(i, x, j)}{\eta(i, j)}$$

kde $\eta(i, x, j)$ je počet všetkých najkratších ciest medzi vrcholmi v_i a v_j, ktoré prechádzajú vrcholom v_x a $\eta(i, j)$ celkový počet najkratších ciest medzi vrcholmi v_i a v_j.

Pri výpočte centrality blízkosti, harmonickej centrality a centrality medzipolohy je nutné v kontexte s korelačnými sieťami pracovať s maticou vzdialeností definovanou vzťahom (1).

Centralitu vrcholu môžeme definovať aj v zmysle jeho susedstva. Môžeme predpokladať, že nejaký vrchol je dôležitý práve vtedy ak je spojený s inými dôležitými vrcholmi v sieti. Ak dôležitosť kvantifikujeme vektorom y, potom môžeme definovať dôležitosť vrcholu v_x ako

$$y_{v_x} = \frac{1}{\lambda}\sum_{i=1}^{|V|} a_{xi} y_{v_i}$$

kde $\lambda \neq 0$ je konštanta. V matice forme vieme vzťah prepísať do tvaru

$$\lambda y = Ay$$

Vidíme, že dôležitosť y nie je nič iné ako vlastný vektor matice susednosti. Prvky vektora y sa nazývajú centralitami vlastného vektora. Vlastné číslo λ je vybrané ako maximálne vlastné číslo. Navyše ak prvky matice A sú nezáporné, vektor y má násobnosť 1 a jeho prvky majú rovnaké znamienko, teda ich vieme preškálovať na kladné. Nevýhodou tejto centrality je, vzhľadom na štruktúru siete, že najväčšia časť centralít je koncentrovaná do pár vrcholov. V takom prípade, väčšine vrcholov bude pridelená centralita blízka nule.

Poslednou centralitou, ktorú si uvedieme, je centralita PageRank. Tá je v princípe založená na náhodnej prechádzke sieťou. Definuje sa ako

$$\pi^T = \pi^T G$$

kde G je takzvaná Google matica definovaná ako

$$G = \kappa\left(P + \frac{ae^T}{|V|}\right) + \left(\frac{1-\kappa}{|V|}\right) ee^T$$

κ je tlmiaci faktor, a je vektor osamostatnených vrcholov, ktorého prvky sú buď 1 ak je vrchol osamostatnený a 0 ak nie je. Pod pojmom osamostatnený vrchol si predstavujeme vrchol, z ktorého nevychádza žiadna hrana. $|V|$ je počet vrcholov v sieti a e je jednotkový vektor dĺžky $|V|$. Matica P je matica pravdepodobností prechodu z jedného vrcholu do druhého

$$P_{ij} = \frac{a_{ij}}{\sum_j a_{ij}}$$

kde P_{ij} predstavuje pravdepodobnosť prechodu z vrcholu v_i do vrcholu v_j. PageRank centralita sa dá chápať ako pravdepodobnosť, že sa nakoniec dostaneme do vrcholu v_i po veľkom množstve krokov náhodnej prechádzky po sieti.

Pre výpočet centrality vlastného vektora aj centrality PageRank je nutné pracovať s kladnými váhami hrán sieti. Preto pred výpočtom oboch centralít z korelačnej siete transformujeme váhy hrán pomocou logistickej transformácie

$$\overline{w}_{ij} = \frac{exp(w_{ij})}{1 + exp(w_{ij})}$$

Globálnu mieru centrality môžeme vyjadriť pomocou miery centralizácie [27]. Tá má veľmi jednoduchý predpis a môže byť počítaná na základe ľubovoľne zvolenej centrality vrcholov.

$$Cent(G) = \sum_{i=1}^{|V|} (max(c(v), v) - c(v_i), v_i)$$

Táto miera nám hovorí o tom, o koľko viac centrálny je najcentrálnejší vrchol sieti oproti všetkým ostatným vrcholov v sieti.

2 Predikcia hrán

Predikcia hrán v sieťach, ako už bolo spomenuté v úvode, je jednou z aktuálnych tém. Za posledné dve dekády vzniklo veľa prístupov založených na rôznych predpokladoch. V krátkosti prejdeme základné rozdelenie metód určených na predikciu hrán. Jeden zo smerov si v práci osvojíme, zovšeobecníme a aplikujeme ho na korelačné siete, pričom bude podrobne zdôvodnené prečo sme si ho vybrali a ako jeho zovšeobecnenie pre daný typ sietí má vyzerať.

V podstate sa dajú metódy predikcie hrán rozdeliť do troch väčších skupín, pričom každá skupina je zostavená z viacerých podkategórií.

2.1 Prehľad rôznych prístupov

2.1.1 Metódy založené na podobnosti vrcholov

Prvú veľkú skupinou metód na predikciu hrán budú tvoriť prístupy založené na podobnosti vrcholov, ktoré využívajú topologické vlastnosti siete. Sú veľmi priamočiare a patria k najstarším metrikám na predikciu hrán. Daná skupina metód sa snaží predikovať chýbajúce alebo budúce spojenia, pričom priraďuje každej dvojici vrcholov skóre podobnosti podľa určitých štrukturálnych vlastností siete. Čím viac sú samotné vrcholy v nejakom zmysle podobné, tým je pravdepodobnosť, že budú spojené hranou vyššia. Samotná skupina sa dá ešte rozdeliť do troch menších skupín podľa veľkosti topológie, ktorú využívajú.

Prvú podkategóriu tvoria metódy založené na lokálnych topologických vlastnostiach. Hlavnou myšlienkou je predpoklad, že ak dvojica vrcholov má podobné susedstvo v rámci siete, tak pravdepodobne sformujú spojenie niekedy v budúcnosti. Zároveň treba podotknúť, že celá skupina takýchto prístupov je veľmi časovo priateľská pri implementácií. Túto kategóriu ale nebudeme nateraz bližšie rozvádzať, nakoľko predstavuje jadro tejto práce a vrátime sa k nej v podkapitole 2.2.

Druhú kategóriu tvoria metódy využívajúce globálne topologické vlastnosti. V porovnaní s lokálnymi metódami využívajú celú topológiu siete, nie len susedstvá jednotlivých vrcholov. Výhodou je, že nie sú limitované v priraďovaní podobnosti pre dvojice vrcholov, ktoré sú od seba vzdialené v sieti a nemajú ani jedného spoločného

suseda. Čím ale daná skupina stráca, je časová náročnosť, ktorá je o poznanie vyššia ako pri lokálnych metódach. Vo väčšine prípadov sú metódy z tejto skupiny založené na ceste medzi vrcholmi. Ako príklad si uvedieme Katzov index podobnosti [28], globálny Leicht-Holme-Newmanov index [29], SimRank [30] alebo pseudo-inverzia laplacovej matice [31].

Poslednú podkategóriu tvoria kvázilokálne metódy. Tie využívajú rozšírené susedstvo každého vrcholu, teda nie len priamych susedov. Vznikli ako trade-off medzi časovou nenáročnosťou, ktorú ponúkajú lokálne metódy a rozšírenou informáciou, ktorú zase ponúkajú globálne metódy. Ponúkajú možnosť, ako pomerne časovo lacno spočítať podobnosť vrcholov, ktoré nemajú spoločných susedov ale stále sú pomerne blízko seba v pohľade na celú sieť. Rovnako si uvedieme zopár príkladov, ktorými sú index lokálnej cesty [32], lokálna náhodná prechádzka [33] alebo index FriendLink [34].

2.1.2 Pravdepodobnostné prístupy

Pravdepodobnostné modely sa dajú považovať za modely učenia s učiteľom, ktoré využívajú pomerne rozsiahlo Bayesovo pravidlo. Ich najväčšou nevýhodou je časová a výpočtová náročnosť, hlavne pre väčšie siete. Patria sem hierarchické štrukturálne modely [35], založené na pozorovaniach, že veľa skutočných sietí vykazuje štrukturálnu topológiu, stochastické blokové modely [36], ktoré sú založené na predpoklade, že vrcholy, ktoré sú medzi sebou rozsiahlo poprepájané, by mali formovať blok alebo komunitu alebo aj pravdepodobnostný model zovšeobecneného koeficientu zhlukovania [37], ktorý, na rozdiel od pôvodného koeficientu zhlukovania, pracuje s cyklami veľkosti k.

2.1.3 Metódy strojového učenia

Ich hlavnou myšlienkou je vyjadriť sieť pomocou matice čŕt, to znemená, že pomocou rôznych algoritmov sa snažia každému vrcholu priradiť vektor charakteristických čŕt. Následne sa dá potom úloha predikcie hrán chápať ako klasifikačná úloha, respektíve regresná úloha ak pracujeme s váženými sieťami. Zavedenie týchto modelov viedlo k zvýšeniu presnosti pri predikovaní hrán [39]. Tieto prístupy vieme kategorizovať do troch kategórií podľa toho, akú metódu používajú na

mapovanie siete, ako merajú podobnosť vrcholov a zavedenú stratovú funkciu. Patria sem metódy maticovej faktorizácie, ktoré sú schopné extrahovať aj nepozorované črty zo sieti. Poväčšine pracujú s maticou podobností, pričom sa snažia o redukciu dimenzie ale zároveň o zachovanie nelineárnych vzťahov a lokálnych štruktúr [40], [41]. Metódy využívajúce prechádzky po sieti, ktoré zostavujú črty a podobnosti vrcholov preskúmavaním siete rôznymi algoritmami preskúmavania [42], [43] a metódy využívajúce konvolučné neurónové siete [44], [45], [46].

2.2 Lokálne metódy a ich zovšeobecnenie

Cieľom tejto práce je, okrem samotného zovšeobecnenia metód na predikciu hrán, aj vytvorenie algoritmu na filtrovanie informácie. Ten budeme neskôr v práci vytvárať na postupnom pridávaní hrán do určitej formy minimálnej kostry. Z toho dôvodu je pre nás dôležité, aby dané metódy zbiehali veľmi rýchlo. Preto iné metódy, ako metódy využívajúce lokálnu topológiu siete, nie sú pre naše použitie vhodné.

2.2.1 Spoloční susedia

Metrika spoloční susedia (CN) je jednou z najpoužívanejších a zároveň aj najefektívnejších metrík, napriek svojej jednoduchosti [47]. Myšlienka za ňou je veľmi intuitívna, a to taká, že pravdepodobnosť, že dva vrcholy sformujú spojenie medzi sebou je priamo ovplyvnená počtom ich spoločných susedov. Inak povedané, dva vrcholy s vysokou pravdepodobnosťou vytvoria hranu medzi sebou, ak je počet ich spoločných susedov vysoký. CN je definovaný ako

$$CN(v_x, v_y) = \left|\Gamma(v_x) \cap \Gamma(v_y)\right|$$

kde $\Gamma(*)$ opäť reprezentuje množinu všetkých susedov daného vrcholu. Všeobecne pre všetky nasledujúce metriky bude myšlienka ich zovšeobecnenia pre znamienkové vážené siete založená na tom, aby bola zachovaná pôvodná myšlienka indexu, ale aby zároveň brala do úvahy aj váhy aj znamienka váh jednotlivých hrán. V literatúre možno nájsť zovšeobecnenie lokálnych indexov podobnosti pre vážené siete [48], avšak ako za okamih demonštrujeme, v našom prípade to nebude postačujúce. Skôr ako sa dostaneme k zovšeobecneniu daného indexu si zavedieme označenia, ktoré sa nám budú hodiť kvôli estetickým úpravám a prehľadností jednotlivých indexov.

Pre ľubovoľnú trojicu vrcholov v_x, v_a, v_y budeme symbolom $w_{x,a,y}$ označovať súčin

$$w_{x,a,y} = w_{xa}w_{ya}(|w_{xa}| + |w_{ya}|)$$

symbolom $\Gamma(v_x, +)$ a $\Gamma(v_x, -)$ budeme označovať množiny takých susedov vrcholu v_x, ktoré sú s týmto vrcholom spojené hranou s kladnou váhou, respektíve zápornou.

Index spoloční susedia sa pre vážené siete používa v tvare [48]

$$WCN(v_x, v_y) = \sum_{v_a \in \Gamma(v_x) \cap \Gamma(v_y)} w_{xa} + w_{ya}$$

čo však v našom prípade nie je postačujúce. Ukážeme si to na jednoduchom prípade. Zoberme si trojicu vrcholov v_x, v_a, v_y s hranami $w_{xa} = -0{,}7$ a $w_{ya} = -0{,}3$. Podľa daného predpisu má index vážených spoloční susedia hodnotu -1. To by nám vyjadrovalo, že dané dva vrcholy sú nepodobné, čo je v rozpore s intuíciou, ktorá nám hovorí, že oba vrcholy v_x a v_y sú s vrcholom v_a spojené hranou s rovnakým znamienkom, a teda by mali byť podobné. Prirodzene sa nám núka využiť absolútnu hodnotu súčtu váh hrán. To čiastočne rieši náš problém, keďže hodnota indexu by v prípade $w_{xa} = -0{,}7$ a $w_{ya} = -0{,}3$ mala hodnotu 1. Čo je však problémom pri uvažovaní absolútnej hodnoty je to, že vzniká problém, ak hrany nemajú rovnaké znamienko. Uvažujme tento krát okrajové hodnoty pre korelačný koeficient $w_{xa} = -1$ a $w_{ya} = 1$. Obyčajný súčet, ale aj absolútna hodnota nám dávajú hodnotu indexu 0, čo opäť je trošku neintuitívne, pretože by sme čakali, že takéto dva vrcholy budú maximálne nepodobné. Zároveň ak by sme zobrali druhý extrémny prípad a zobrali veľmi malé čísla, napríklad $w_{xa} = -0{,}1$ a $w_{ya} = 0{,}1$, hodnota indexu s absolútnou hodnotou by bolo opäť 0. To by dávalo síce väčší zmysel z pohľadu intuície, že o ich podobnosti veľa povedať nevieme, pretože sú obe spojené pomerne slabo s vrcholom v_a, avšak problém spočíva v tom, že by sme chceli takéto prípady jednoznačne od seba odlíšiť. Z praktického hľadiska sa ukazuje ako rozumné riešenie teda nezobrať súčet, ako je to v prípade zovšeobecnenia pre vážené siete, ale zobrať súčin váh. Tým sa opäť čiastočne vyriešia problémy ako súčtu váh, tak absolútnej hodnoty súčtu. V prípade $w_{xa} = -1$ a $w_{ya} = 1$by sme dostali hodnotu indexu -1, čo by bolo v súlade s intuíciou, v prípade $w_{xa} = -0{,}1$ a $w_{ya} = 0{,}1$ by sme dostali hodnotu indexu $-0{,}01$, čo je tiež fajn. Opäť ale nastáva mierny problém s nejednoznačnosťou v určitých prípadoch. Zoberme

si dve dvojice váh w_{xa} a w_{ya}. Prvá bude mať hodnoty $w_{xa} = 1$ a $w_{ya} = 0{,}5$, čím dostávame hodnotu indexu so súčinom váh 0,5. Druhú dvojicu si zvolíme $w_{xa} = 0{,}8$ a $w_{ya} = 0{,}625$, čím opäť dôjdeme k hodnote 0,5. Pritom je zrejme, že prvá dvojica je silnejšia ako druhá, keďže súčet váh v prvom prípade je 1,5, zatiaľ čo v druhom iba 1,425. Keďže aj takéto situácie by sme chceli, použijeme súčin definovaný pomocou $w_{x,a,y}$, ktoré je vyjadrené na začiatku podkapitoly. Poslednou úpravou, ktorá bude spoločná pre všetky indexy bude namiesto čisto súčinu $w_{x,a,y}$, spravíme geometrický priemer súčiniteľov, a to z dôvodu aby výsledná hodnota súčinu bola aspoň rádovo rovnaká ako hodnoty váh. To znamená, že aj výsledný index spoloční susedia pre znamienkové vážene siete bude mať podobu

$$SWCN(v_x, v_y) = \sum_{v_a \in \Gamma(v_x) \cap \Gamma(v_y)} \left(w_{x,a,y}\right)^{1/3}$$

Alternatívou je namiesto geometrické priemeru zvoliť zovšeobecnený geometrický priemer, ktorý by mal tvar

$$(w_{xa})^{\alpha_1}\left(w_{ya}\right)^{\alpha_2}\left(|w_{xa}| + |w_{ya}|\right)^{\alpha_3}$$

pričom $\alpha_1 + \alpha_2 + \alpha_3 = 1$.

Vo všeobecnosti, hlavne ak daná sieť nie je korelačnou sieťou, sa dá postupovať inak. Napríklad sa zavedie namiesto mocniteľa $1/3$ premenná $\alpha \in (0, \infty)$, ktorej funkcia bude buď nadhodnocovať malé váhy a podhodnocovať veľké, a to tak, že zvolíme $\alpha \in (1, \infty)$ alebo naopak, budeme chcieť malé váhy podhodnotiť a veľké váhy nadhodnotiť, a to tak, že zvolíme $\alpha \in (0{,}1)$, čo je aj náš prípad.

Iný postup, ktorý by viedol k lepšej aproximácií váženého indexu, aj v prípade, že by váhy hrán boli v absolútnej hodnote väčšie ako 1, spočíva v násobený súčiniteľa $\left(|w_{xa}| + |w_{ya}|\right)$ funkciou s argumentom $w_{xa} * w_{ya}$. Ak by sme pracovali s korelačnými sieťami, hodnoty váh sú v absolútnej hodnote menšie ako 1, postačujúcou funkciou by bola mocninová funkcia s nepárnym veľkým stupňom mocniny α

$$w_{x,a,y} = f\left(w_{xa} * w_{ya}\right) * \left(|w_{xa}| + |w_{ya}|\right)$$

$$f\left(w_{xa} * w_{ya}\right) = \left(w_{xa} * w_{ya}\right)^{\alpha}$$

Ľahko sa dá vidieť, ak pošleme α do nekonečna (stále platí, že α je nepárne číslo) je $f\left(w_{xa} * w_{ya}\right) = -1$, ak $w_{xa} * w_{ya} < 0$, $f\left(w_{xa} * w_{ya}\right) = 1$, ak $w_{xa} * w_{ya} > 0$ a $f\left(w_{xa} * w_{ya}\right) = 0$, ak $w_{xa} * w_{ya} = 0$.

Ak by sme chceli pracovať s hodnotami váh v absolútnej hodnote väčšími ako 1, vhodnou funkciou pre $f(w_{xa} * w_{ya})$ by bola škálovaná funkcia hyperbolického tangensu

$$f(w_{xa} * w_{ya}) = \tanh\left(\frac{w_{xa} * w_{ya}}{n}\right).$$

Väčšina nasledujúcich indexov bude založených principiálne rovnako, ako je index spoloční susedia s tým rozdielom, že sa zavedú určité penalizácie.

2.2.2 Jaccardov index

Pre binárnu sieť je Jaccardov index definovaný veľmi podobne ako index spoloční susedia, s tým rozdielom, že celkový počet spoločných susedov je normalizovaný celkovým počtom susedov jednotlivých vrcholov [49]. Pre nevážené siete ho teda vieme definovať ako

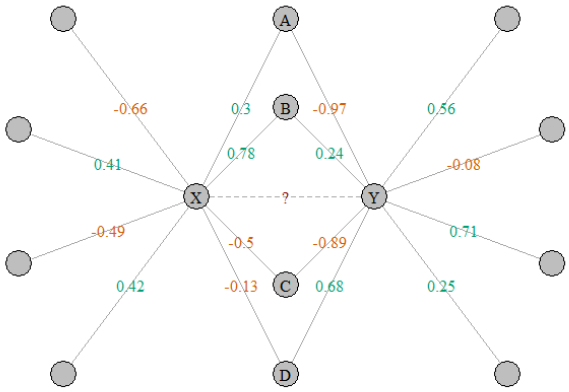
$$JC(v_x, v_y) = \frac{|\Gamma(v_x) \cap \Gamma(v_y)|}{|\Gamma(v_x) \cup \Gamma(v_y)|}$$

Jaccardov index sa dá interpretovať ako percentuálna zhoda susedstiev dvoch vrcholov. Alternatívne sa dá na neho pozerať ako na penalizáciu každého spoločného suseda počtom všetkých vrcholov, ktoré susedia buď s jedným vrcholom alebo druhým. A práve túto interpretáciu budeme chcieť zachovať. Rovnako budeme penalizovať súčin $w_{x,a,y}$, podľa toho aké znamienko majú váhy w_{xa} a w_{ya}. Všetky štyri možnosti dvojíc znamienok si odvodíme teraz s tým, že rovnaká logika sa prenesie aj na zvyšné indexy a ich penalizácie. Ak w_{xa} a w_{ya} majú obe kladné znamienka, potom bude penalizáciu tvoriť iba časť susedstva jednotlivých vrcholov, ktoré sú s danými vrcholmi spojené rovnako kladnou hranou. Obdobne ak w_{xa} a w_{ya} majú obe záporné znamienka, tak v tom prípade bude penalizáciu tvoriť iba časť susedstva jednotlivých vrcholov, a to tých, ktoré sú buď s jedným alebo druhým prepojené zápornou hranou. Ak w_{xa} bude mať kladné znamienko a w_{ya} bude mať záporné znamienko, potom budú penalizáciu tvoriť vrcholy zo susedstva v_x, také, ktoré sú s vrcholom v_x prepojené kladnou hranou a vrcholy zo susedstva v_y, ktoré sú s vrcholom v_y prepojené zápornou hranou. Posledný prípad je analogický, len sa vymenia v_x a v_y. Iný pohľad, ako sa na daný zovšeobecnený index dá pozerať je, že ak si zoberieme ľubovoľného spoločného suseda dvoch vrcholov, tak je penalizovaný silou tých vrcholov z jednotlivých

susedstiev, ktoré sú s danými vrcholmi spojené hranou s rovnakým znamienkom. Ide o akési porovnanie spoločného suseda so susedmi vrcholov, ktorí sú v rovnakom vzťahu s danými vrcholmi (buď sú podobné alebo protichodné).

Podoba Jaccardovho indexu pre znamienkové vážené siete je daná

$$SWJC(v_x, v_y) = \sum_{v_a \in \Gamma(v_x) \cap \Gamma(v_y)} \frac{(w_{x,a,y})^{1/3}}{\sum_{v_b \in \Gamma(v_x, sing(w_{xa}))} |w_{xb}| + \sum_{v_b \in \Gamma(v_y, sing(w_{ya}))} |w_{yb}|}$$



Obrázok 11 Pomocný obrázok pre odvodenie penalizácie JC, SI, LHN, LAS, HD a HP indexu

2.2.3 Saltonov index

Saltonov index sa niekedy označuje aj ako kosínusový index [50]. Dôvodom je, že pre binárne siete je zhodný s kosínusom uhla vektorov matice susednosti. Kosínusová podobnosť definovaná ako kosínus uhla matice susednosti je dobre definovaná aj v prípade znamienkovaných vážených sietí. Preto Saltonov index použijeme iba ako motiváciu na zostavenie podobného indexu pre znamienkové vážené siete. Pre binárne siete je Saltonov index definovaný vzťahom

$$SI(v_x, v_y) = \frac{|\Gamma(v_x) \cap \Gamma(v_y)|}{\sqrt{|\Gamma(v_x)| * |\Gamma(v_y)|}}$$

čo sa teda v maticovej forme dá prepísať ako

$$SI(v_x, v_y) = \frac{A_{x.}^T * A_{y.}}{\|A_{x.}\| * \|A_{y.}\|}$$

kde $A_{i.}$ predstavuje í-ty stĺpec matice susednosti A. Teda Saltonov index je skutočne zhodný s kosínusom uhla, ktoré tieto dva vektory tvoria. My teda nebudeme priamo zovšeobecňovať kosínus uhla, ale vytvoríme nový index, ktorý bude veľmi podobný Jaccardovmu indexu s tým rozdielom, že použijeme inú formu penalizácie. Zovšeobecnený Saltonov index bude mať výslednú podobu

$$SWSI(v_x, v_y) = \sum_{v_a \in \Gamma(v_x) \cap \Gamma(v_y)} \frac{(w_{x,a,y})^{1/3}}{\sqrt{\sum_{v_b \in \Gamma(v_x, sing(w_{xa}))} |w_{xb}| * \sum_{v_b \in \Gamma(v_y, sing(w_{ya}))} |w_{yb}|}}$$

2.2.4 Hub promoted index

Hub promoted index bol pôvodne navrhnutý posudzovanie podobnosti substrátov v metabolických sieťach [51]. Vyjadruje pomer medzi prienikom susedstiev dvoch vrcholov a menšieho zo susedstiev týchto vrcholov. Opäť si to ale vieme interpretovať ako penalizovanie spoločných susedov pomocou veľkosti menšieho zo susedstiev. Matematicky sa definuje ako

$$HP(v_x, v_y) = \frac{|\Gamma(v_x) \cap \Gamma(v_y)|}{min(|\Gamma(v_x)|, |\Gamma(v_y)|)}$$

Zovšeobecnenie teda bude predstavovať penalizáciu jednotlivých spoločných susedov podľa sily menšieho zo susedstiev, pričom sila susedstva sa bude počítať iba podľa tých hrán, ktoré majú rovnaké znamienko ako hrana medzi vrcholom a spoločným susedom.

$$SWHP(v_x, v_y) = \sum_{v_a \in \Gamma(v_x) \cap \Gamma(v_y)} \frac{(w_{x,a,y})^{1/3}}{min\left(\sum_{v_b \in \Gamma(v_x, sing(w_{xa}))} |w_{xb}|, \sum_{v_b \in \Gamma(v_y, sing(w_{ya}))} |w_{yb}|\right)}$$

2.2.5 Hub depressed index

Hub depressed index je opačnou analógiou Hub promoted indexu [52]. V tomto prípade index udáva pomer medzi počtom spoločných susedov a maxima z veľkosti jednotlivých susedstiev daných vrcholov. Matematicky je definovaný nasledovne

$$HD(v_x, v_y) = \frac{|\Gamma(v_x) \cap \Gamma(v_y)|}{max(|\Gamma(v_x)|, |\Gamma(v_y)|)}$$

Jeho zovšeobecnenie bude vyzerať analogicky ako zovšeobecnenie hub promoted indexu.

$$SWHD(v_x, v_y) = \sum_{v_a \in \Gamma(v_x) \cap \Gamma(v_y)} \frac{(w_{x,a,y})^{1/3}}{max\left(\sum_{v_b \in \Gamma(v_x, sing(w_{xa}))} |w_{xb}|, \sum_{v_b \in \Gamma(v_y, sing(w_{ya}))} |w_{yb}|\right)}$$

2.2.6 Leicht-Holme-Newmanov index

LHN index je veľmi podobný Saltonovmu indexu, jediný rozdiel je v tom, že menovateľ zo Saltonovho indexu v tomto prípade nie je odmocnený [53]. Vyjadriť sa dá v tvare

$$LHN(v_x, v_y) = \frac{|\Gamma(v_x) \cap \Gamma(v_y)|}{|\Gamma(v_x)| * |\Gamma(v_y)|}$$

pričom zovšeobecnenie je úplne analogické ako v prípade Saltonovho indexu

$$SWLHN(v_x, v_y) = \sum_{v_a \in \Gamma(v_x) \cap \Gamma(v_y)} \frac{(w_{x,a,y})^{1/3}}{\sum_{v_b \in \Gamma(v_x, sing(w_{xa}))} |w_{xb}| * \sum_{v_b \in \Gamma(v_y, sing(w_{ya}))} |w_{yb}|}$$

2.2.7 Local affinity structure index

LAS index zobrazuje príbuznosť vzťahov medzi dvojicou vrcholov a ich spoločnými známymi. Opiera sa o hypotézu, že čím je vyššia táto príbuznosť dvoch vrcholov a ich susedstva, tým je väčšia šanca, že dané dva vrcholy budú v budúcnosti prepojené hranou [54]. Opäť si to môžeme predstaviť ako penalizáciu spoločných susedov podľa toho, ako veľké sú jednotlivé susedstvá.

$$LAS(v_x, v_y) = \frac{|\Gamma(v_x) \cap \Gamma(v_y)|}{|\Gamma(v_x)|} + \frac{|\Gamma(v_x) \cap \Gamma(v_y)|}{|\Gamma(v_y)|}$$

Analogicky tento index zovšeobecníme ako predošlé indexy do tvaru

$$SWLAS(v_x, v_y) =$$

$$= \sum_{v_a \in \Gamma(v_x) \cap \Gamma(v_y)} \left(\frac{(w_{x,a,y})^{\frac{1}{3}}}{\sum_{v_b \in \Gamma(v_x, sing(w_{xa}))} |w_{xb}|} + \frac{(w_{x,a,y})^{\frac{1}{3}}}{\sum_{v_b \in \Gamma(v_y, sing(w_{ya}))} |w_{yb}|} \right)$$

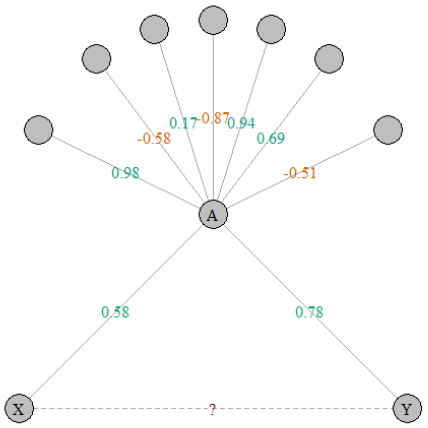
2.2.8 Adamic-Adarov index

Pôvodný účel metriky bolo porovnanie dvoch webových stránok medzi sebou [55]. Jednoducho pracuje s myšlienkou dať väčšiu váhu susedom s menšími susedstvami. Podoba Adamic-Adarovho indexu je nasledovná

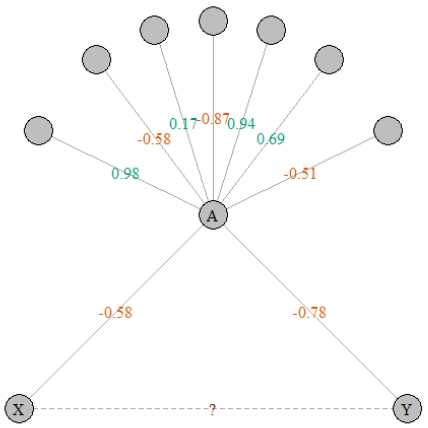
$$AA(v_x, v_y) = \sum_{v_a \in \Gamma(v_x) \cap \Gamma(v_y)} \frac{1}{ln|\Gamma(v_a)|}$$

Na rozdiel od predošlých indexov, táto metrika pracuje okrem susedstiev dvoch vrcholov, ktorých podobnosť sa snažíme odhadnúť, aj so susedstvom každého spoločného suseda oboch vrcholov. Myšlienka AA indexu sa dá chápať aj nasledovne. Ak by sme mali v sieti silný vrchol s veľa spojeniami, tak pravdepodobnosť, že tento vrchol sa objaví v nejakom spoločnom susedstve je veľká. V takomto ponímaní sa dajú chápať ako dôležitejšie vrcholy, ktoré majú málo spojení v rámci siete, ale stále sú spoločným susedom oboch vrcholov. V prípade zovšeobecnenia chceme zachovať právu túto úvahu. Znamená to, že daná penalizácia bude silne závisieť opäť od znamienka hrany, ktorá spája daného suseda s jedným alebo druhým vrcholom. Teda ak by boli obe hrany záporné, do penalizácie bude vstupovať iba susedstvo vrcholu, ktoré sú s daným vrcholom spojené kladnou hranou. Čo nám bude hovoriť o tom, aká veľká je pravdepodobnosť, že je spoločným susedom dvoch vrcholov vzhľadom na počet kladných hrán vychádzajúcich z neho. V prípade, že obe hrany sú záporné bude myšlienka analogická, iba sa bude brať ohľad na susedstvo vzhľadom na záporné váhy hrán. Ak by nastala možnosť, že jedna hrana je kladná a druhá záporná, do úvahy sa bude brať celé susedstvo daného vrcholu. Nakoniec aby sme teda vyrovnali penalizácie pri všetkých štyroch možnostiach, v prípade rôznych znamienok hrán budeme dané susedstvo brať polovične, respektíve s polovičnou silou. Finálna podoba zovšeobecneného AA indexu bude mať tvar

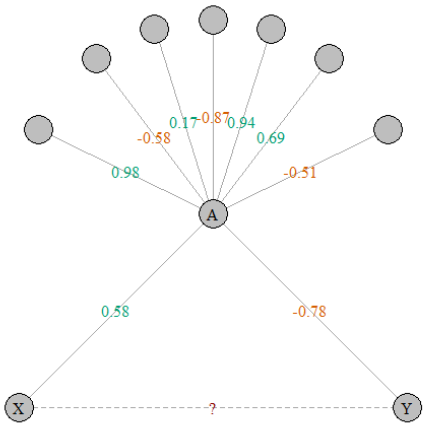
$$SWAA(v_x, v_y) = \sum_{v_a \in \Gamma(v_x) \cap \Gamma(v_y)} \frac{(w_{x,a,y})^{1/3}}{ln\left(1 + \frac{\sum_{v_b \in \Gamma(v_a, sing(w_{xa}))} |w_{ab}| + \sum_{v_b \in \Gamma(v_a, sing(w_{ya}))} |w_{ab}|}{2}\right)}$$



Obrázok 12 Pomocný obrázok pre odvodenie penalizácie AA a RA indexu



Obrázok 13 Pomocný obrázok pre odvodenie penalizácie AA a RA indexu



Obrázok 14 Pomocný obrázok pre odvodenie penalizácie AA a RA indexu

2.2.9 Resource Allocation Index

Veľmi podobným idenxom Adamic-Adarovho indexu je práve Resource allocation index [56]. Ich rozdiel nastáva len v tom, že zatiaľ čo AA index penalizuje spoločných susedov podľa logaritmu veľkosti ich susedstva, RA index logaritmus vypúšťa.

$$RA(v_x, v_y) = \sum_{v_a \in \Gamma(v_x) \cap \Gamma(v_y)} \frac{1}{|\Gamma(v_a)|}$$

Zovšeobecnenie je úplne analogické tomu pre AA index s rozdielom logaritmu.

$$SWRA(v_x, v_y) = \sum_{v_a \in \Gamma(v_x) \cap \Gamma(v_y)} \frac{(w_{x,a,y})^{1/3}}{\left(\frac{\sum_{v_b \in \Gamma(v_a, sing(w_{xa}))} |w_{ab}| + \sum_{v_b \in \Gamma(v_a, sing(w_{ya}))} |w_{ab}|}{2}\right)}$$

Zostávajúce dva indexy, ktoré ešte budeme používať v práci, budú špecifické tým, že ich nebudeme zovšeobecňovať ale budeme používať podobu, akú majú. Reč je o indexe Preferential attachment [57], ktorej základná myšlienka znie, že nový vrchol pridaný do siete sa s najväčšou pravdepodobnosťou spojí s najsilnejším vrcholom. Formulovať ho môžeme nasledovne

$$PA(v_x, v_y) = |\Gamma(v_x)| * |\Gamma(v_y)|$$

pričom jeho zovšeobecnenie [48] nám dáva zmysel aj pre znamienkové vážené siete.

$$SWPA(v_x, v_y) = \sum_{v_a \in \Gamma(v_x))} w_{xa} * \sum_{v_b \in \Gamma(v_y)} w_{yb}$$

Samotný index, či už pôvodný alebo zovšeobecnený, sa nepozerá na spoločných susedov. Postačujúci pre nás je z dôvodu, že ak u oboch vrcholov prevládajú, respektíve sú silnejšie hrany rovnakých znamienok, potom aj výsledná hodnota indexu bude mať kladné znamienko a bude hovoriť o ich podobnosti. Ak by u jedného vrcholu boli silnejšie kladné prepojenia a u druhého záporné, potom sú si v istom zmysle tie dva vrcholy nepodobné, čo sa prejaví a výsledná hodnota indexu bude mať záporné znamienko. Preto v tomto prípade netreba robiť ďalšie úpravy a index sa dá použiť tak, ako bol odvodený.

Posledným indexom podobnosti je kosínusová podobnosť, ktorá, ako bolo spomínané vyššie, je dobre definovaná pre ľubovoľnú maticu susednosti a rovnako môže byť použitá bez nášho zásahu.

2.3 Analýza fungovania zovšeobecnených indexov

V tejto kapitole demonštrujeme použitie indexov na filtrovaných sieťach, ktoré dostaneme vytvorením minimálnej kostry, TMFG grafu a na plne prepojené kompletnej sieti. Porovnáme medzi sebou správanie sa indexov a pozrieme sa aj nato, či sa im darí oddeliť hraný, ktoré by mali byť kladné od tých, ktoré by mali byť záporné. Ako prvé si v krátkosti predstavíme dáta, s ktorými budeme pracovať.

2.3.1 Dáta

Dáta, s ktorými budeme aj po zvyšok práce pracovať, predstavujú hodnoty stovky európskych akcií, ktoré majú najvýraznejšie percentuálne zastúpenie v európskom akciovom indexe STOXX EUROPE 600. Tento index tvorí skoro 600 európskych akcií s rôznymi váhami. My budeme pracovať len so sto akciami s najvyššími váhami v indexe. Dáta boli stiahnuté z voľne dostupného zdroja https://finance.yahoo.com/ pomocou R-kového balíčka QUANTMOD [38] pomocou funkcie getSymbol. Hodnoty pochádzajú z obdobia medzi rokmi 2016 až 2021,

a predstavujú denné hodnoty akcií. Z výnosov akcií vypočítame maticu Pearsonových korelačných koeficientov, ktoré budú tvoriť sieť, s ktorou budeme celý čas pracovať.

Predtým, ako pristúpime k samotnému porovnávaniu indexov, je potreba povedať jednu vec. Keďže aj minimálna kostra aj TMFG by s veľkou pravdepodobnosťou vytvorili filtrované siete, ktoré by boli z väčšej časti, pri minimálnej kostre aj z celej časti, tvorené iba kladnými hranami, rozhodli sme sa filtrované siete vytvárať z matice absolútnych hodnôt korelačných koeficient. Dôvodom je, že chceme porovnať ich fungovanie aj pri zastúpení záporných hrán, a teda záporné a kladné korelácie budeme brať s rovnakou váhou. Inými slovami informáciu, ktorú nám dávajú záporné korelácie budeme chápať ako veľkosť protichodnosti. Algoritmus pre minimálnu kostru a TMFG nám určia v akomsi zmysle dôležité hrany v takomto pojímaní korelácií, pričom priamo do filtrovaných sietí zvolíme tieto hrany s pôvodnými hodnotami, teda aj záporné.

Keďže väčšina indexov je založená na využití spoločných susedov, je zrejmé, že pre dvojice vrcholov, ktoré nebude mať ani jedného spoločného suseda bude hodnota každého indexu 0, okrem indexu PA, ktorý spoločných susedov nepoužíva. Z tohto dôvodu budeme pri porovnávaní indexov brať do úvahy iba ich hodnoty pre dvojice, ktoré majú aspoň jedného suseda. Rovnako nás nebudú zaujímať hodnoty medzi dvojicami, ktoré už sú priamo prepojené hranou vo filtrovanej sieti, okrem časti kde sa pozrieme na podobnosť vrcholov v rámci plnej siete. Tu prirodzene nemá zmysel pozerať na počet spoločných susedov, rovnako nemá zmysel uvažovať hrany, ktoré ešte nie sú v sieti pridané.

Indexy budeme analyzovať medzi sebou na základe hodnoty korelačného koeficientu medzi hodnotami indexov pre dvojice, ktoré sme určili vyššie. Na porovnanie použijeme tri rôzne korelačné koeficienty. Prvým bude klasický najviac využívaný Pearsonov korelačný koeficient [58]. Pri predikcií hrán nás ani tak nezaujíma veľkosť hodnôt samotných indexov, skôr chceme zistiť, ktorá hrana ma najväčšiu šancu vzniknúť. Budeme chcieť zisťovať ako veľmi sa dané indexy zhodujú pri zoradení hrán podľa ich pravdepodobnosti budúceho vzniku. Na to využijeme Spearmanov korelačný koeficient [59] a Kendallovo Tau [60].

Majme dvojice hodnôt (x_1, y_1),..., (x_n, y_n)

- Pearsonov korelačný koeficient

$$\rho_{x,y}^{pears} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$\bar{x}$ a $\bar{y}$ sú jednotlivé výberové priemery.

- Spearmanov korelačný koeficient

Označme si zoradené hodnoty (x, y) ako $rx_i = rank(x_i), ry_i = rank(y_i)$. Potom Spearmanov korelačný koeficient sa dá definovať ako

$$\rho_{x,y}^{spear} = \rho_{rx,ry}^{pears} = \frac{\sum_{i=1}^{n}(rx_i - \overline{rx})(ry_i - \overline{ry})}{\sqrt{\sum_{i=1}^{n}(rx_i - \overline{rx})^2}\sqrt{\sum_{i=1}^{n}(ry_i - \overline{ry})^2}}$$

Ak žiadne dva ranky sa nezhodujú, t. j. nemáme rovnaké hodnoty, potom sa alternatívne dá Spearmanov korelačný koeficient vyjadriť ako

$$\rho_{x,y}^{spear} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

$$d_i = rank(x_i) - rank(y_i)$$

- Kendallovo Tau

$$\tau_{x,y} = \frac{2}{n(n^2 - 1)} \sum_{i<j} \text{sgn}\left(x_i - x_j\right)\text{sgn}\left(y_i - y_j\right)$$

Rovnako ako Spearmanov korelačný koeficient, aj Kendallovo Tau meria koreláciu medzi usporiadaním jednotlivých postupností. Vyjadruje rozdiel medzi počtom dvojíc pozorovaní (x_i, y_i) a $\left(x_j, y_j\right), i < j$, pre ktoré platí $x_i < x_j \wedge y_i < y_j$ alebo $x_i > x_j \wedge y_i > y_j$, to znamená, že pri oboch pozorovaniach sa zhoduje usporiadanie a počtom dvojíc pozorovaní (x_i, y_i) a $\left(x_j, y_j\right), i < j$, pre ktoré platí $x_i < x_j \vee y_i > y_j$ alebo $x_i > x_j \vee y_i < y_j$, to znamená, že usporiadanie rovnaké nemajú.

Ako posledné budeme porovnávať, či sa zhodujú predpoklady aj na znamienko budúcej hrany s hodnotou indexu. Inými slovami, chceme zistiť, ak majú niektoré hrany, ktoré chceme pridávať, v pôvodnej sieti kladné znamienko a niektoré záporné, či ich indexy vedia od seba dostatočne odlíšiť. Nato použijeme vizualizačné metódy, ktoré nám to pomôžu veľmi rýchlo zistiť. Konkrétne použijeme metódu hlavných komponentov [61] a následne vykreslíme hodnoty indexov pre dané hrany v dvoj rozmernom priestore.

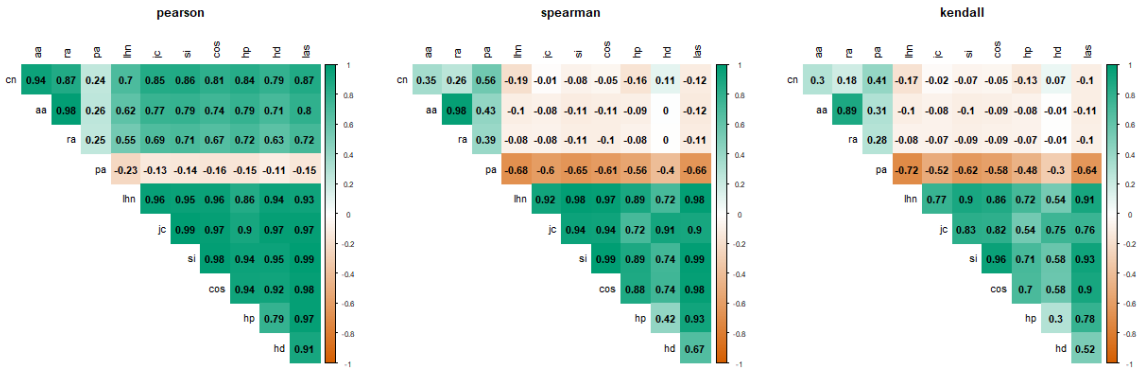
Keďže metóda hlavných komponentov nie je hlavná časť práce, prejdeme si len základnú myšlienku a vzťahy. Tou je zredukovať dimenziu dát aby sme zachovali čo najväčšiu informáciu z nich. Metóda hlavných komponentov pracuje s kovariančnou

maticou a nie je invariantná na škálovanie. Z tohto dôvodu je v praxi lepšie pracovať s korelačnou maticou. Informáciou budeme chápať ako rozptýlenosť dát, čiže hľadáme smer najväčšieho rozptylu dát. Hlavné komponenty sú nové premenné, ktoré sú skonštruované ako lineárna kombinácia pôvodných premenných. Sú vytvorené tak, aby boli nové premenné nekorelované a väčšina informácie bola stlačená do prvých komponentov, tak aby prvý komponent niesol najväčšiu časť, druhý druhú najväčšiu a tak ďalej. Metóda hlavných komponentov úzko súvisí s rozkladom korelačnej matice pomocou vlastných vektorov a vlastných čísel. Keďže sa jedná o symetrickú kladne semidefinitnú maticu, jednotlivé vlastné vektory sú na seba kolmé a vlastné čísla udávajú ich natiahnutie v danom smere. Práve najväčšie vlastné číslo prislúcha vlastnému vektoru, v ktorom sú dáta najviac rozptýlené. Pre náš prípad zoberieme prvé dva smery najväčšieho rozptylu a dáta projektujeme do priestoru, ktorý tvoria.

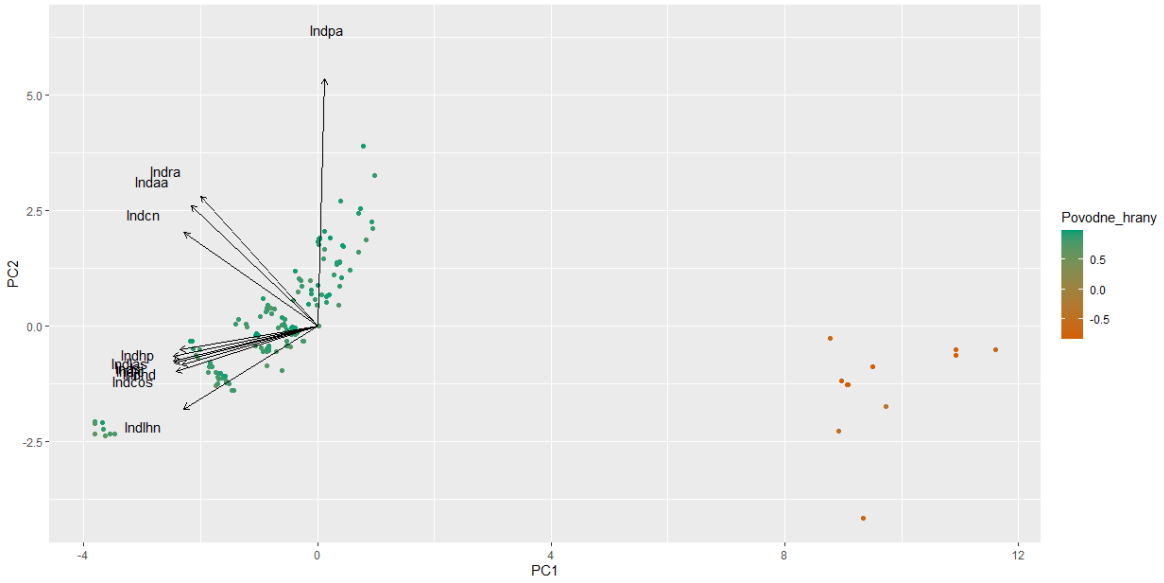
2.4 Porovnávanie indexov

Začneme analýzou indexov v prípade najmenšej siete, teda minimálnej kostry. Až na PA index možno vidieť, že hodnoty Pearsonových korelačných koeficientov sú dosť vysoké. Dôvodom je, že hodnoty samotných indexov sa menia približne rovnako, respektíve rozdiely vzhľadom na hodnoty sú malé. Na druhej strane ak sa pozrieme ako zoraďujú hrany podľa pravdepodobnosti, že vzniknú, môžeme si všimnúť delenie indexov na skupinky. Jednu skupinu tvoria indexy PA a CN. Druhú skupinu tvoria indexy AA a RA, ktoré majú veľmi podobný predpis, líšia sa minimálne. Tretiu skupinu tvoria zvyšné indexy, ktoré majú opäť veľmi podobný predpis, obrázok 15.

Vzhľadom na vykreslenie indexov v dvoj rozmernom priestore môžeme konštatovať, že vedia veľmi dobre rozlíšiť aj to hrana s akým znamienkom sa očakáva. Evidentne sú rozlíšiteľné dva zhluky. Jeden, ktorý tvoria kladné hrany a druhý, ktorý tvoria záporné hrany, obrázok 16.



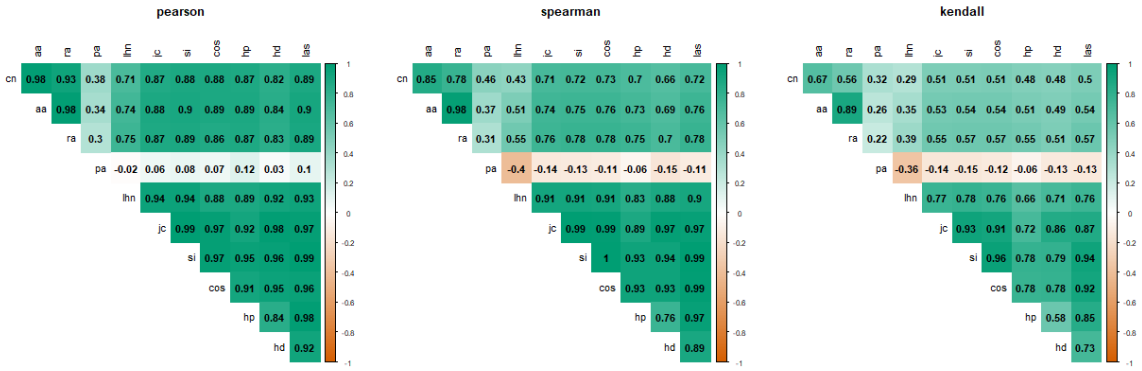
Obrázok 15 Matica korelačných koeficientov indexov – minimálna kostra grafu



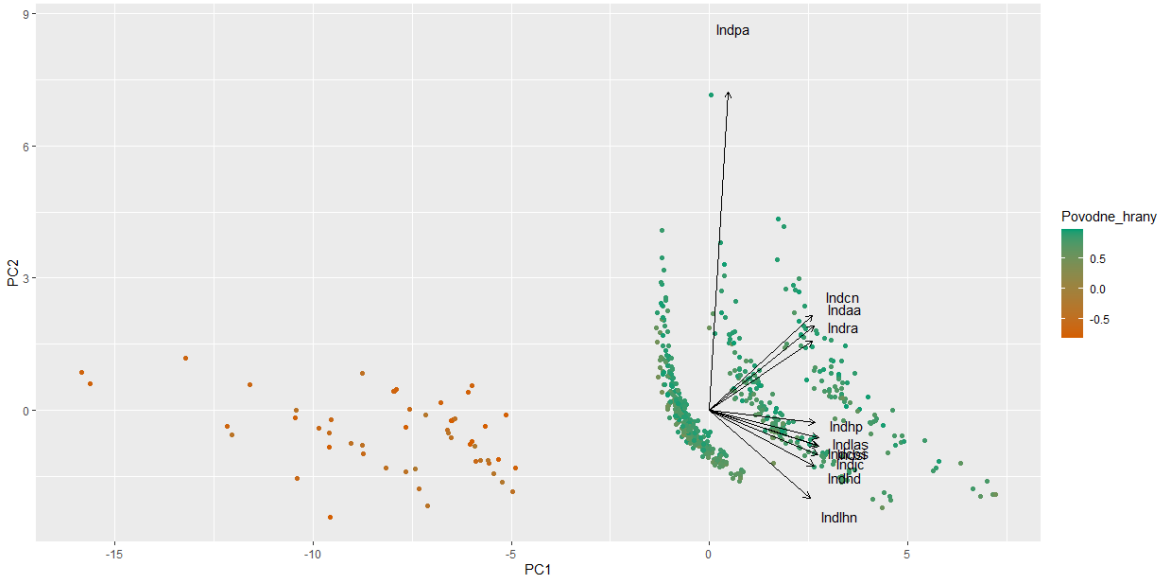
Obrázok 16 Vykreslenie v dvojrozmernom priestore pomocou metódy hlavných komponentov – minimálna kostra grafu

V prípade TMFG siete dostávame iné výsledky. Matice Pearsonových korelačných koeficient je síce veľmi podobná, no matica Spearmanových korelačných koeficientov a Kendallových tau koeficientov naznačuje značné zvýšenie podobnosti, obrázok 17. Chápať sa to dá aj tak, že vo filtrovanej sieti sú dvojice vrcholov, ktoré si doslova pýtajú byť spojené hranou. Dôvodom je samotný algoritmus, ktorý sieť vytvára, keďže je obmedzený v pridávaním hrán a často sa môže stať, že dva vrcholy sú si veľmi podobné, no kvôli porušeniu podmienok hrana nemôže byť pridaná.

Čo sa týka rozlíšenia toho aké znamienko by mala mať potencionálne pridaná hrana, vedia indexy dobre určiť aj v prípade TMFG siete. Potenciálnych hrán je v tomto prípade viacej, a je vidieť náznak prechodu sprava doľava vzhľadom na znamienka hrán. Vidíme, že úplne napravo sú striktne kladné hrany, v stredu sú to hrany s hodnotami blízkymi nule a naľavo sú to striktne záporné hrany, obrázok 18.

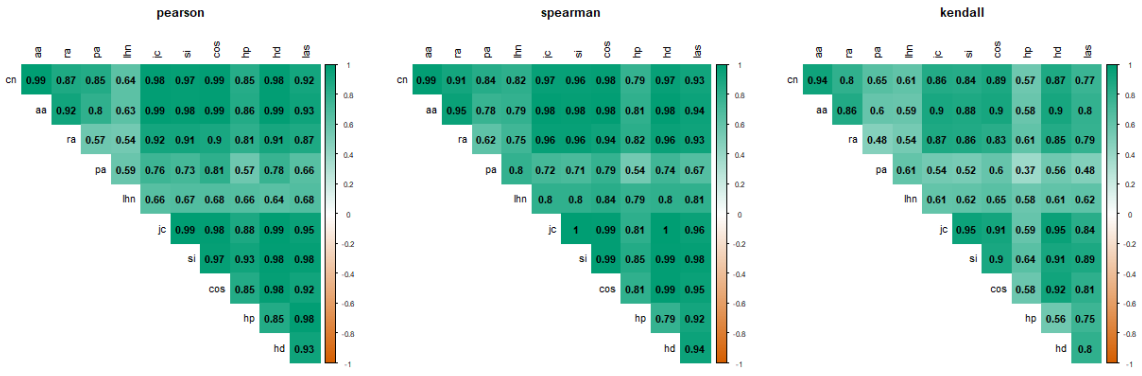


Obrázok 17 Matica korelačných koeficientov indexov – TMFG

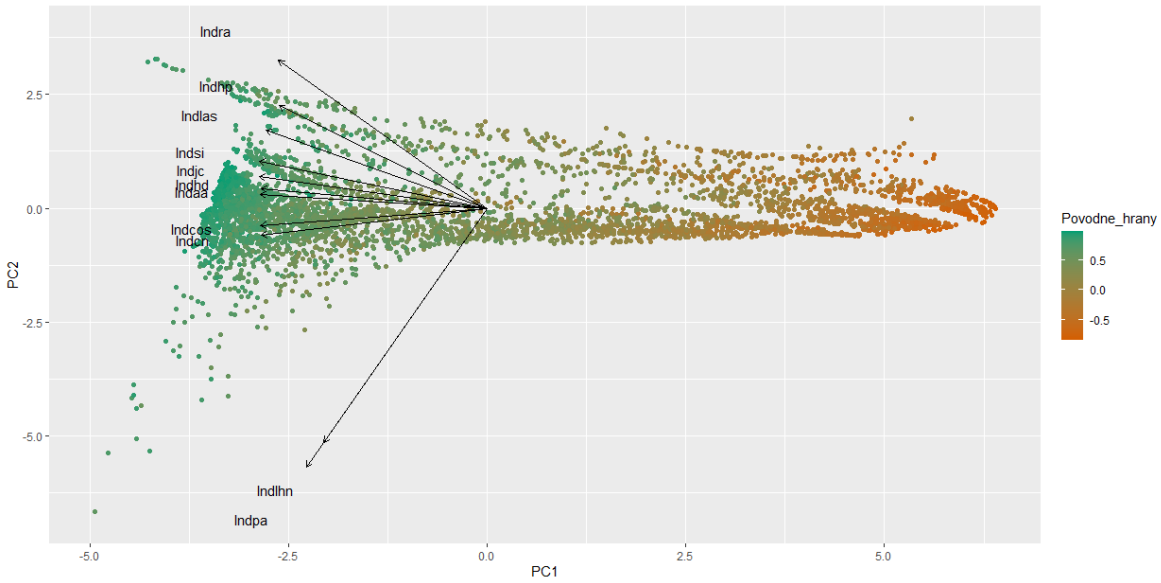


Obrázok 18 Vykreslenie v dvojrozmernom priestore pomocou metódy hlavných komponentov – TMFG

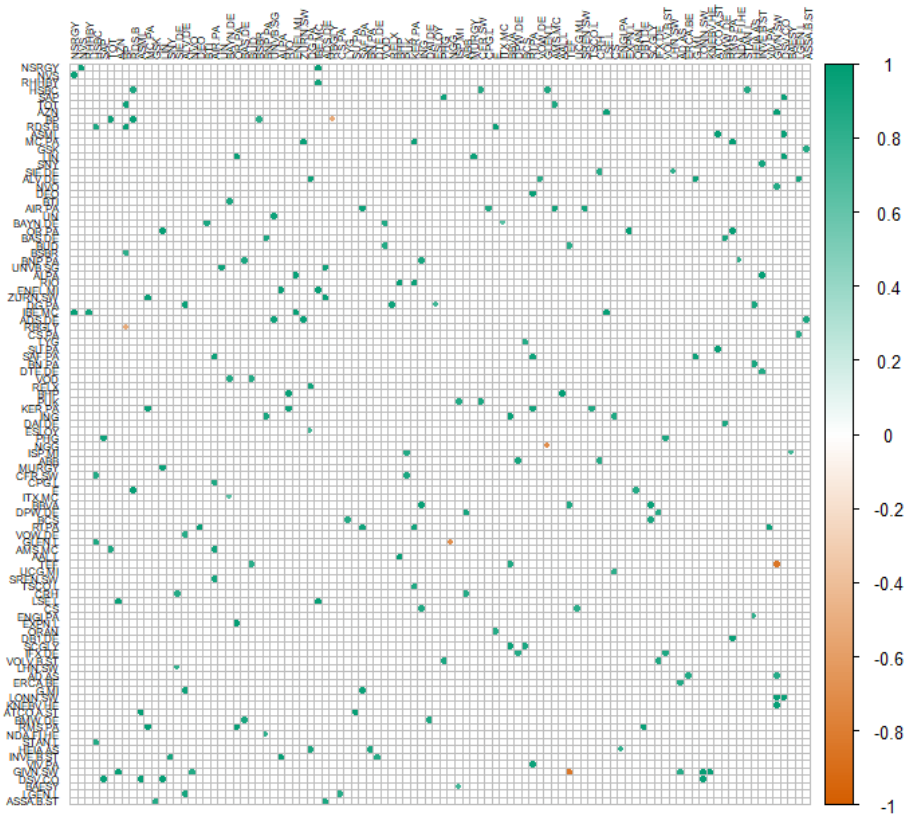
Ako posledné sa pozrieme, ako sa správajú indexy vzhľadom na plnú sieť. Môžeme vidieť veľmi očakávaný výsledok, a tým je, že sú si všetky indexy nesmierne podobné, obrázok 19. Dôvodom je, že pri plnej sieti sú susedstvá jednotlivých vrcholov skoro totožné. Keď sa pozrieme na predpis všetkých indexov, vidíme, že v takomto prípade sú jednak rovnaké menovatele jednotlivých zlomkov a viac menej sú rovnaké aj čitatele a presne preto nie je žiadnym prekvapením, že v rámci plnej siete sú si indexy tak veľmi podobné. Jediná otázka, ktorú treba v tomto prípade zodpovedať je, či sa viac menej zhodujú pozorované hrany s dosiahnutou podobnosťou. V tomto prípade vidíme už plynulý prechod od kladných k záporným hranám, okrem pár výnimiek, ktoré tvoria hrany v strede, obrázok 20. To je ale pochopiteľné, keďže hodnoty týchto hrán sú blízke nule. Pravdepodobne môžeme pozorovať to, že hoci medzi dvoma vrcholmi je malá kladná hrana, podľa topológie a hodnôt ich zvyšných spojení by sme možno skôr očakávali malú zápornú hodnotu hrany. Rozdiel v tomto prípade je ale zanedbateľný, hlavne z dôvodu, že pri metóde hlavných komponentov strácame určitú časť informácie. Tá by mohla, ale nemusela, v konečnom výsledku dobre od seba rozlíšiť aj tieto veľmi malé hrany. Keďže ani v jednom prípade by sa nestalo, že by výrazne kladné a výrazne záporné hrany boli pomiešané, môžeme skonštatovať, že indexy v týchto príkladoch fungujú tak, ako by sme od nich očakávali.



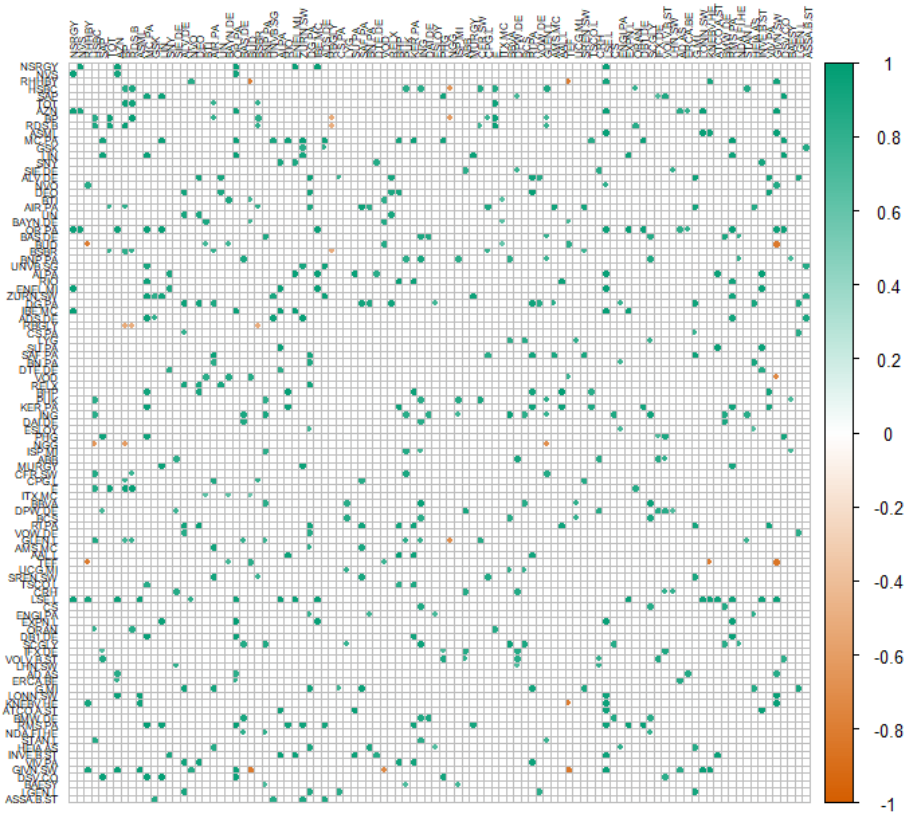
Obrázok 19 Matica korelačných koeficientov indexov – plná sieť



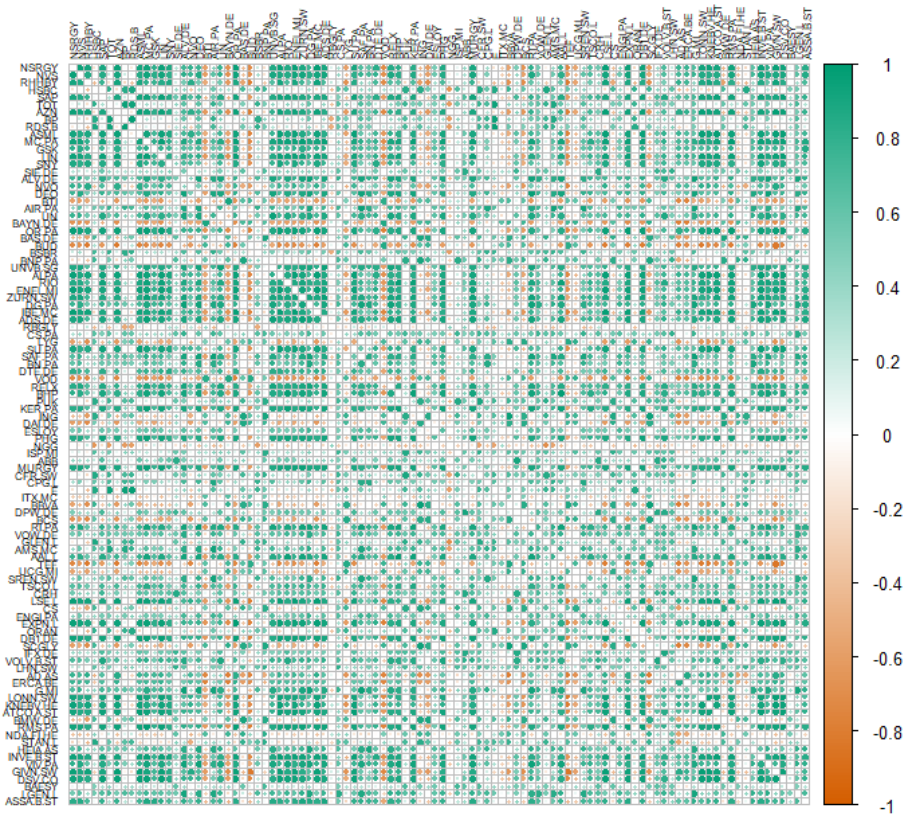
Obrázok 20 Vykreslenie v dvojrozmernom priestore pomocou metódy hlavných komponentov – plná sieť



Obrázok 21 Matica susednosti – minimálna kostra grafu



Obrázok 22 Matica susednosti – TMFG



Obrázok 23 Matica susednosti – plná sieť

3 Filtrovanie informácie zo siete

Ako bolo naznačené v kapitole 1, pracovať s plnou sieťou nemusí byť vždy priamočiare a jednoduché. Dôležitá informácia môže byť často prekrytá a prakticky nedostupná kvôli vysokému počtu hrán. Pre zjednodušenie práce sa používajú práve filtrované siete, ktoré nám pomáhajú pochopiť hlavnú štruktúru danej siete. Rovnako už bolo spomenuté, že najčastejšie sa využíva minimálna kostra grafu, planárne maximálne filtrovaný graf alebo jeho aproximácia v podobe trojuholníkovo maximálne filtrovaného grafu. Problémom minimálnej kostry, respektíve jej nevýhodou, je počet hrán rovný $n-1$, kde n predstavuje počet vrcholov. Ak si to porovnáme s počtom všetkých možných hrán siete s n vrcholmi $\frac{n(n-1)}{2}$, vidíme, že maximálny počet hrán je násobne vyšší, ako počet hrán v minimálnej kostre. To znamená, že pri väčších sieťach môže byť informácie z minimálnej kostry veľmi okresaná a podstatná časť hlavnej štruktúry siete môže ostať nevydolovaná. V prípade PMFG je počet hrán vyšší a predstavuje hodnotu $3n-6$. Ako nevýhodu v tomto prípade môžeme chápať podmienku planarity, pre ktorú nie je v podstate dobrý dôvod, prečo práve ona ma zabezpečiť vyfiltrovanie dôležitej časti informácie. Filtrovaná sieť, ktorá vzniká metódou víťaz berie všetko môže byť v konečnom dôsledku nesúvislá, čo by mohlo vyvolať určité problémy. Problémom je často aj zvoliť hranicu, podľa ktorej sa hrany zo siete odstraňujú. Z tohto dôvodu sme sa rozhodli prísť s alternatívnou metódou, ktorá by aplikovala zovšeobecnené indexy podobnosti z kapitoly 2 na zostavenie filtra, ktorý by nám určil dôležitú štruktúru na základe podobností samotných vrcholov. Docieliť by sme chceli plnšiu filtrovanú sieť oproti minimálnej kostre grafu, súvislú sieť na rozdiel od metódy víťaz berie všetko a úplne relaxovať podmienku planarity výslednej siete. Daný algoritmus budeme zostavovať prihliadajúc na jeho uplatnenie vo filtrovaní informácie z korelačnej siete.

3.1 Algoritmus založený na podobnosti vrcholov

Samotný algoritmus bude pozostávať z niekoľko krokov. V prvom rade je dobré uvedomiť si, ako by mal začať. To úzko súvisí s myšlienkou samotného algoritmu. Ťažiskom algoritmu bude v každom kroku pridať do filtrovanej siete jednu hranu, podľa toho, ako podobné sú si vrcholy v rámci danej filtrovanej siete. Znamená to, že namiesto toho, aby sme použili iba pôvodnú sieť, budeme vytvárať akúsi postupnosť podsietí, kde sa dve nasledujúce budú líšiť maximálne o jednu hranu. Následne

budeme skúmať, ako sa vyvíjajú jednotlivé metriky siete, s ktorými sme sa oboznámili v kapitole 1. Ak sa teda vrátime k myšlienke, ako daný algoritmus začať, jedno z riešení, ktoré využijeme my v práci, je začať minimálnou kostrou zo siete vytvorenej pomocou absolútnych hodnôt korelácií, vytvorenej rovnako ako v predošlej kapitole. Tá nám bude určovať iba dôležité hrany, nie ich váhy. Váhy budú stále hodnoty pôvodných korelačných koeficientov.

3.1.1 Algoritmus

Algoritmus podrobne predstavíme v niekoľkých krokoch.

i. Vytvorenie minimálnej kostry zo siete absolútnych hodnôt korelačných koeficientov

Týmto krokom je zabezpečené, že ak sa v pôvodnej sieti nachádzajú dôležité záporné hrany, nebudú prehliadané. Opäť pripomíname, že minimálna kostra bude určovať iba ktoré hrany majú ostať, pričom ich hodnoty budú totožné s pôvodnou korelačnou sieťou. Taktiež si tým zabezpečíme, že hodnota indexov bude môcť dosahovať aj záporné hodnoty a tým zabezpečíme, že aj táto stránka štruktúry bude zohľadnená.

ii. Vypočítanie podobnosti pomocou zvoleného indexu pre dvojice vrcholov

Podobnosť sa vypočíta pre všetky dvojice vrcholov, ktoré nespája hrana v aktuálnej podsieti.

iii. Obrátenie určitých hodnôt indexu

Na určenie najpravdepodobnejšej hrany potrebujeme istým indexom otočiť znamienko. Dôvodom je, že chceme pridávať aj záporne hrany, ktorých dôležitosť bude zakomponovaná v zápornej podobnosti medzi vrcholmi. Preto pre všetky dvojice vrcholov, ktoré sú v pôvodnej sieti spojené zápornou hranou, obrátime znamienko indexu podobnosti.

iv. Pridanie hrany

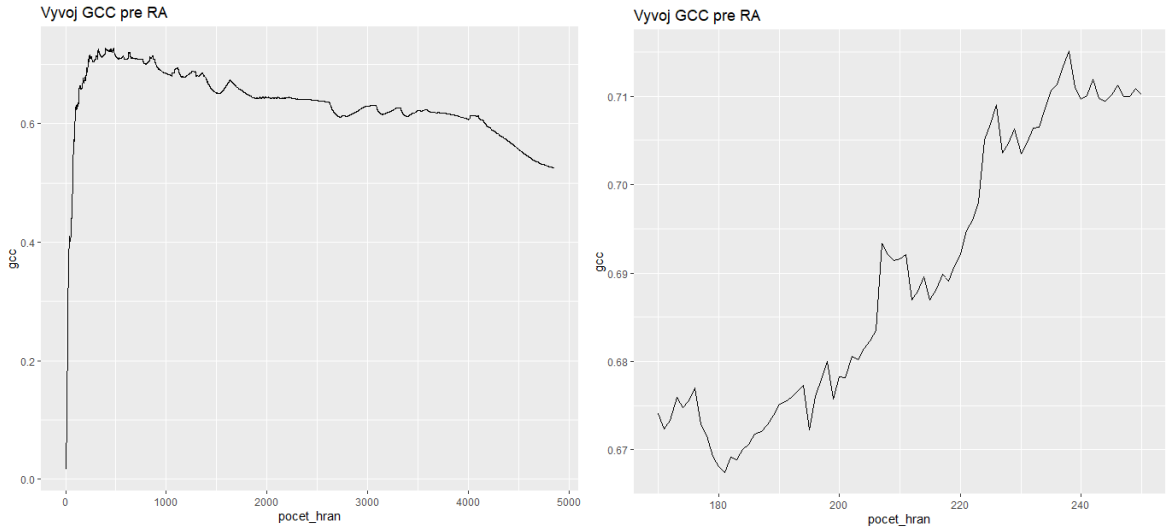
Pridáme hranu medzi dva vrcholy, ktoré sú si najviac podobné podľa daného indexu. Prípad, že by sme pridali hranu opačného znamienka ako má hodnota podobnosti, nemôže nastať, pretože takéto hrany budú mať po obrátení znamienka zápornú podobnosť. Jediný prípad kedy by sa niečo takéto mohlo prihodiť by bol, keby ani jedno znamienko hrany nekorešpondovalo so znamienkom podobnosti.

v. Kroky ii. iii. a iv. opakujeme pokiaľ sa nepridajú všetky hrany alebo neurčíme počet hrán, kedy ma algoritmus skončiť

Tento algoritmus nám vracia postupnosť hrán, ako sú pridávané. Dobrý počet hrán budeme určovať kvalitatívne podľa priebehu globálneho koeficientu zhlukovania a centralizácie. Nasledujúca demonštrácia bude ukázaná na indexe RA, ktorý, ako sa neskôr ukáže, je jedným z najvhodnejších indexov na použitie pri filtrovaní informácie.

Samotný priebeh koeficientu zhlukovania nám nepovie veľa. Čo z neho ale vieme veľmi dobré vyčítať je, aké podsiete vytvára vzhľadom na prepojenosť susedstiev všetkých vrcholov. Na druhú stranu centralizácia nám bude určovať akési cykly. Záleží od zvolenej centralizácie. Nárast hodnoty predstavuje zvýšenie centrálnosti najcentrálnejšieho vrcholu v zmysle danej centrality. Toto bude spravidla prebiehať iba pár pridaní hrán a následne zaznamenáme dlhší pokles, kedy bude stúpať centralita ostatných vrcholov.

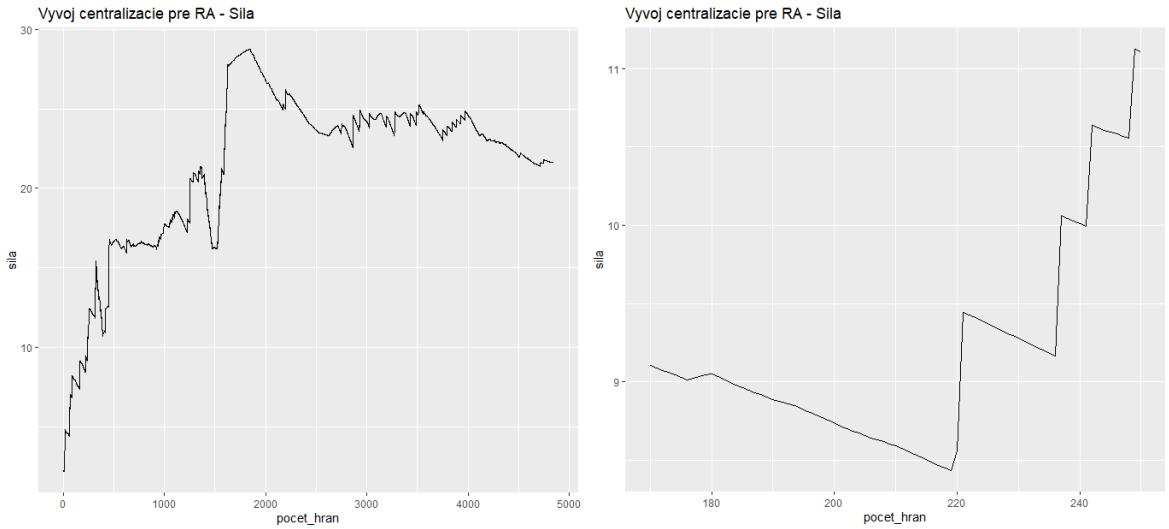
Konkrétne na prípade indexu RA môžeme pozorovať nasledovné.



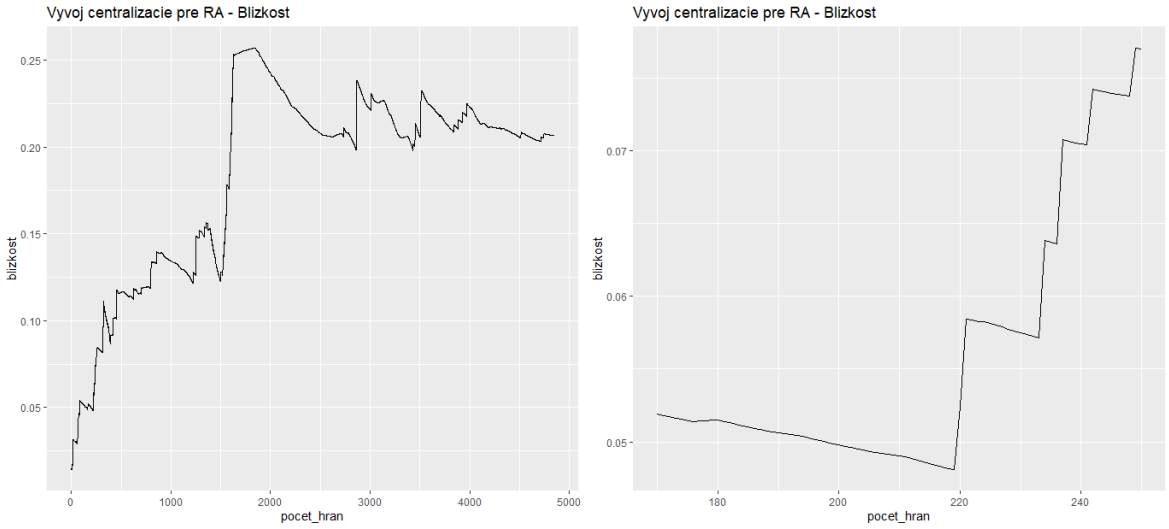
Obrázok 24 Vývoj hodnoty globálneho koeficientu zhlukovania pri pridávaní podľa indexu RA – vpravo celkový priebeh, vľavo priblíženie na počet pridaných hrán v rozpätí 170-250

Vývoj globálneho koeficientu zhlukovania je pomerne strmý s následnou chvíľkovou stagnáciou a zdĺhavým poklesom, obrázok 24. Naznačuje nám to, že dosť rýchlo sa poprepájajú hlavne susedstvá vrcholov, no vzhľadom na maximálnu hodnotu niečo viac ako 0.7, sa nepoprepájajú úplne. Čo budeme požadovať my, je aby počet hrán

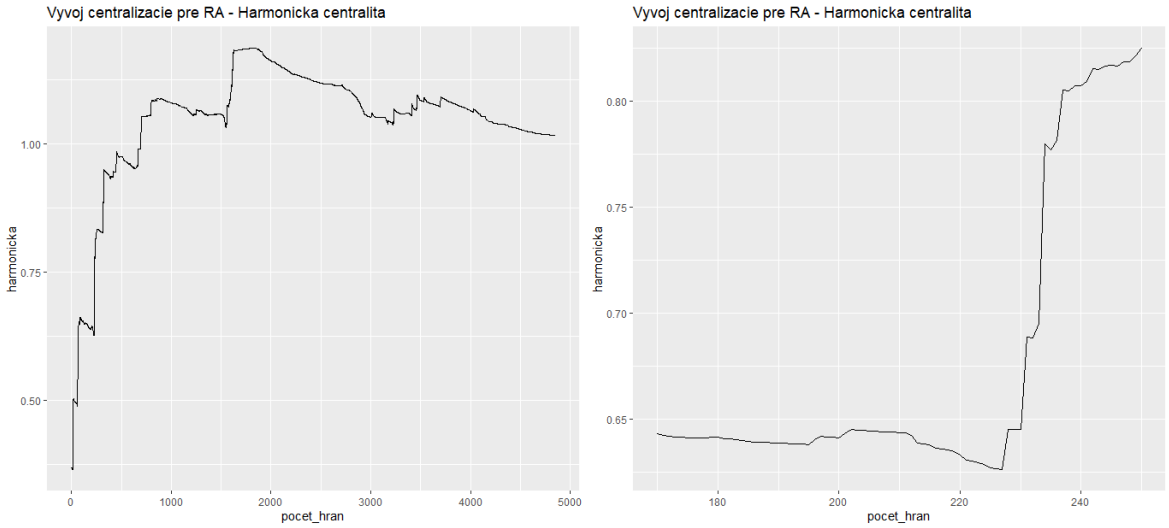
bol niekde v prvotnom stúpaní, ale nie úplne na vrchole. Dôvodom je, že chceme určiť dôležité hrany a nie dostať pár zhlukov úplne prepojených a zhluky spojené iba jednou hranou. Čo sa týka centralizácií, tie sa delia do 2 skupín podobných správaní. Prvú tvoria centralizácie podľa centrality sily, blízkosti, vlastného vektora a PageRank, obrázky 25, 26, 29, 30. Druhú skupinu tvoria centralizácie harmonickej centrality a centrality medzipolohy, obrázky 27 a 28. Hrany budeme určovať v rozpätí $170 - 250$ kvôli podmienke pre globálny koeficient zhlukovania. Taktiež preto, že hrán nechceme pridať príliš veľa aby sme nedostali opäť nepriehľadnú sieť. Ak si priblížime všetky centralizácie v rozsahu $170 - 250$ pridaných hrán, ako dobrý kandidát sa nám javí hodnota 218. Tento bod určuje koniec klesania prvej skupiny centralizácii, zároveň je to bod, v ktorom klesajú centralizácie z druhej skupiny, aj keď to nie koniec klesania. Inak povedané, určuje nám to bod, kedy sa začína ďalšie kolo pridávania hrán k najcentrálnejšiemu vrcholu. Tento bod považujeme za rozumnú hodnotu, kedy algoritmus ukončiť. Podobne budeme postupovať pri zvyšných indexoch, príslušné grafy sa budú nachádzať v prílohe práce.



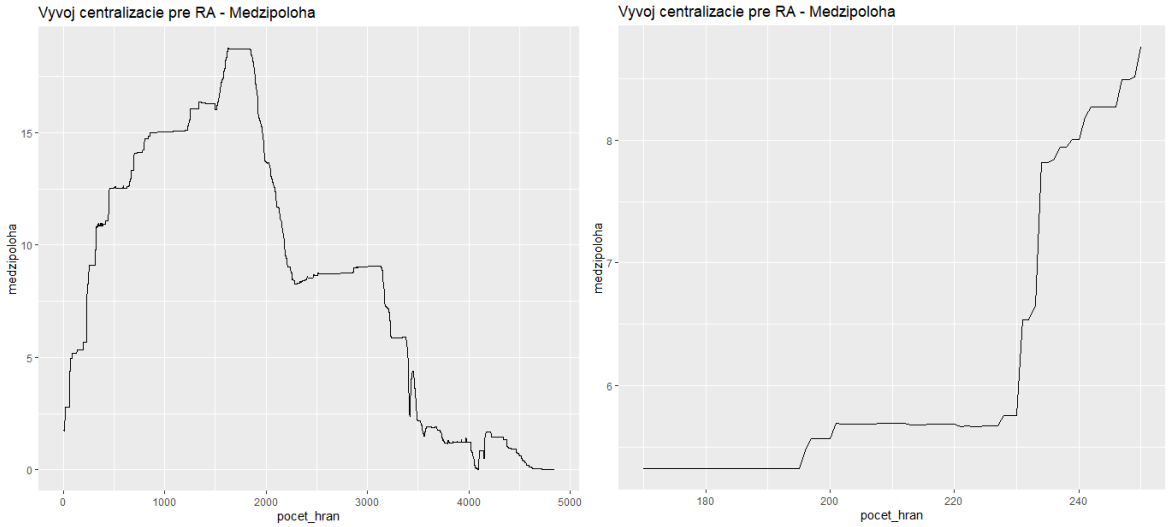
Obrázok 25 Vývoj hodnoty centralizácie sily pri pridávaní podľa indexu RA – vpravo celkový priebeh, vľavo priblíženie na počet pridaných hrán v rozpätí 170-250



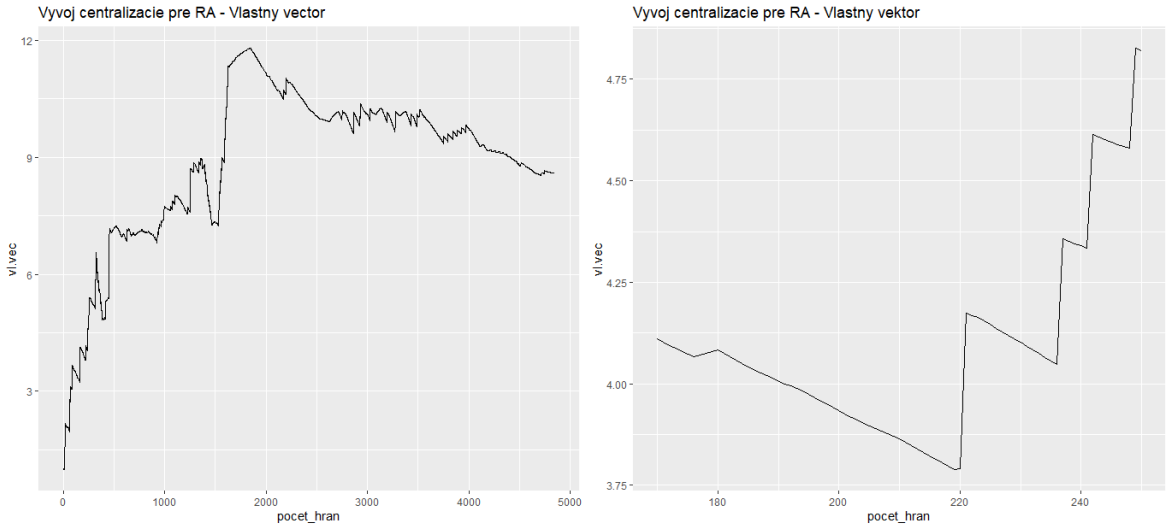
Obrázok 26 Vývoj hodnoty centralizácie blízkosti pri pridávaní podľa indexu RA – vpravo celkový priebeh, vľavo priblíženie na počet pridaných hrán v rozpätí 170-250



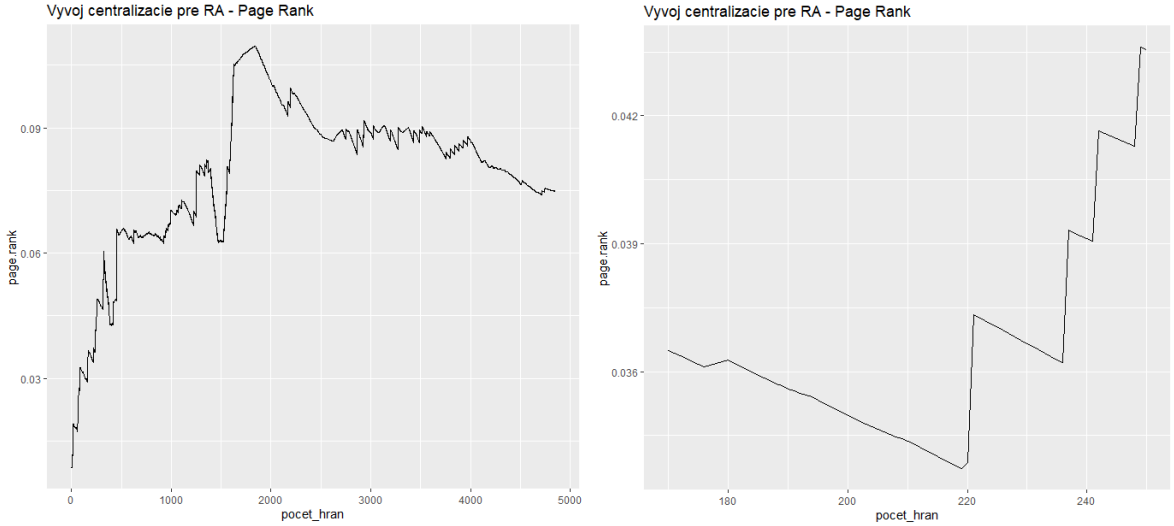
Obrázok 27 Vývoj hodnoty harmonickej centralizácie pri pridávaní podľa indexu RA – vpravo celkový priebeh, vľavo priblíženie na počet pridaných hrán v rozpätí 170-250



Obrázok 28 Vývoj hodnoty centralizácie medipolohy pri pridávaní podľa indexu RA – vpravo celkový priebeh, vľavo priblíženie na počet pridaných hrán v rozpätí 170-250



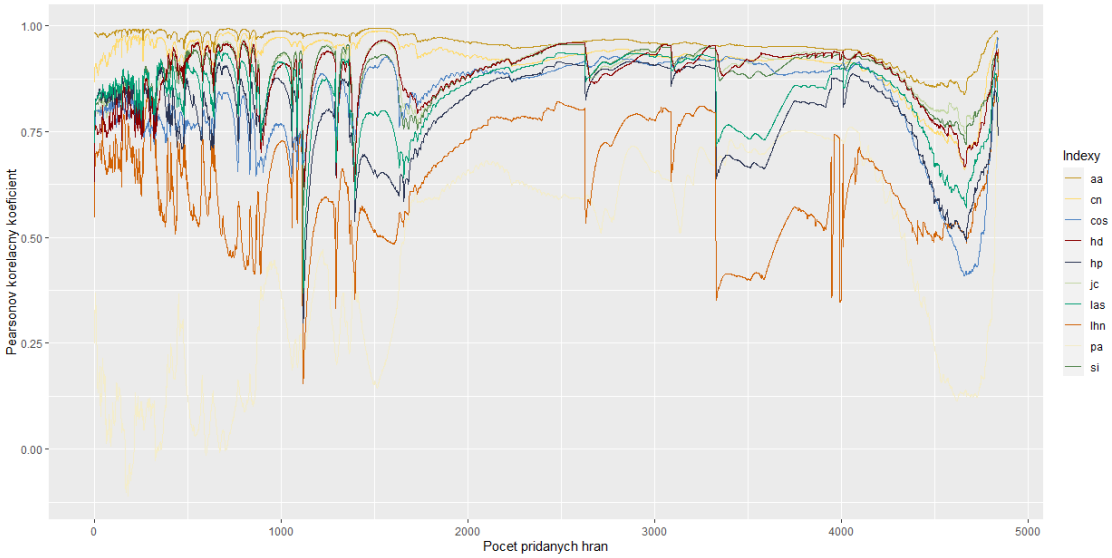
Obrázok 29 Vývoj hodnoty centralizácie vlastného vektora pri pridávaní podľa indexu RA – vpravo celkový priebeh, vľavo priblíženie na počet pridaných hrán v rozpätí 170-250



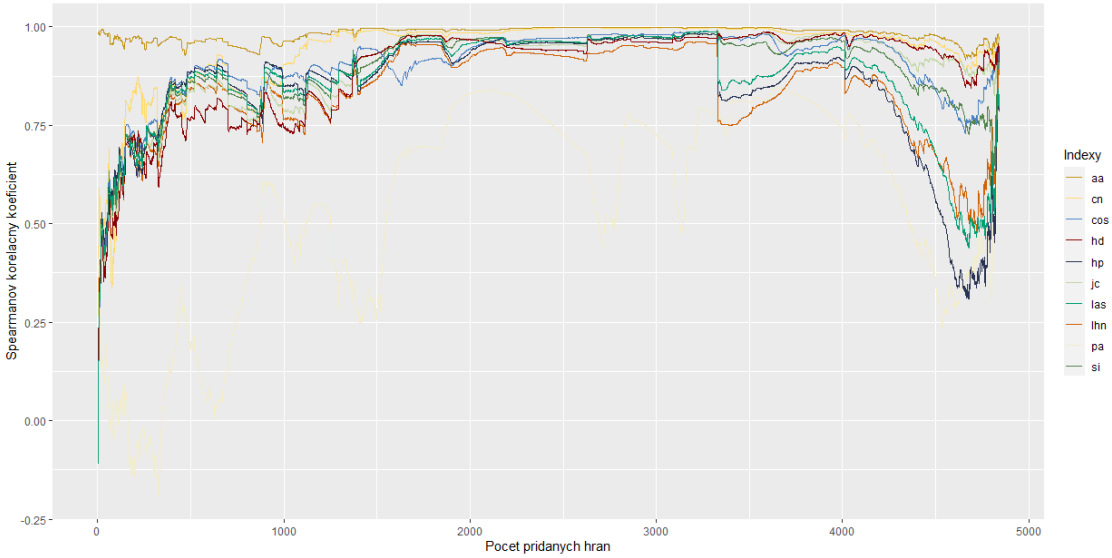
Obrázok 30 Vývoj hodnoty centralizácie PageRank pri pridávaní podľa indexu RA – vpravo celkový priebeh, vľavo priblíženie na počet pridaných hrán v rozpätí 170-250

Na záver ešte porovnáme, ako by sa správali zvyšné indexy na podsieťach vytvorených filtrom založenom na RA indexe v porovnaní s daným indexom RA. Ako aj v predchádzajúcej kapitole využijeme na porovnanie všetky tri druhy korelačných koeficientov. Rovnako to pozorujeme len na dvojiciach vrcholov, ktoré ešte nespája hrana v danej podsieti, ale zároveň majú aspoň jedného suseda. Pearsonov korelačný koeficient vyzerá byť relatívne nestabilný, čo je pochopiteľné. Znamená to iba to, že pridaná hrana rôzne ovplyvňuje hodnoty ostatných koeficientov. Na druhú stranu aj Spearmanov korelačný koeficient aj Kendallovo tau sa javia v zásade stabilne, z čoho

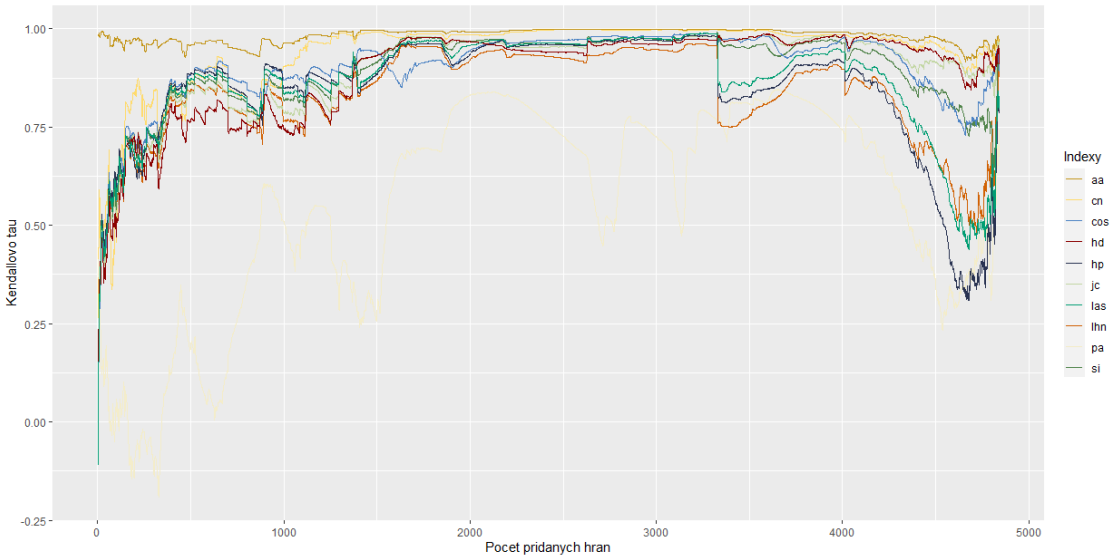
vieme usúdiť napríklad aj to, že pridanie novej hrany, hoc ovplyvňuje výrazne hodnoty indexov, nemení až tak výrazne následného usporiadania hrán, obrázky 31, 32 a 33.



Obrázok 31 Vývoj hodnoty Pearsonovho korelačného koeficientu RA indexu so zvyšnými indexami pri pridávaní podľa indexu RA



Obrázok 32 Vývoj hodnoty Spearmanovho korelačného koeficientu RA indexu so zvyšnými indexami pri pridávaní podľa indexu RA



Obrázok 33 Vývoj hodnoty Kendallovho tau koeficientu RA indexu so zvyšnými indexami pri pridávaní podľa indexu RA

3.2 Filter zohľadňujúci viacero indexov naraz

Prirodzene sa tisne otázka, či zahrnutie viacerých indexov, ba dokonca všetkých, nemôže viesť k zlepšeniu filtra. Existuje mnoho algoritmov určených práve na rozhodovanie sa podľa viacerých kritérií. Hlavnou úlohou týchto algoritmov je určiť preferovanú voľbu z viacerých možností, klasifikovať možností do menšieho počtu kategórii alebo zoradiť možnosti podľa subjektívneho preferovania. Veľmi dobrý a stručný prehľad takýchto algoritmov obsahuje [62].

V práci sa budeme ďalej zaoberať jedným konkrétnym prístupom, a to *Technique for Order of Preference by Similarity to Ideal Solution*, skrátene TOPSIS. Vysvetlime si bližšie fungovanie tohto algoritmu a následne ho využijeme pri filtrovaní informácie zohľadnením všetkých indexov naraz.

3.2.1 TOPSIS algoritmus

TOPSIS algoritmus je jedným z najjednoduchších a zároveň veľmi mocných algoritmov rozhodovania na základe viacerých kritérií [63]. Opiera sa o myšlienku, že vybraná možnosť by mala mať najmenšiu euklidovskú vzdialenosť od potenciálnej najlepšej možnosti a najväčšiu vzdialenosť od potenciálnej najhoršej možnosti. Výrazom potenciálne najlepšia možnosť označíme vektor najlepších hodnôt z každého kritéria, ktoré nemusia teda nutne byť z jednej možnosti. Potenciálne najhoršia možnosť bude ekvivalentne znamenať vektor najhorších hodnôt z každého kritéria,

ktoré opäť nemusia byť nutne z jednej možnosti. Samotný algoritmus pozostáva z nasledujúcich krokov [63]:

Treba spomenúť, že predpokladáme, že najlepšia možnosť pre dané kritérium je tá, ktorá ma najväčšiu hodnotu. Ak by to tak nebolo tak jednoducho pre dané kritérium zameníme maximum minimom pri určovaní najlepšej potenciálnej možnosti, a naopak vymeníme minimum za maximum pri určovaní najhoršej potenciálnej možnosti daného kritéria.

i. Vytvorenie matice alternatív

Riadky matice budú predstavovať rôzne možnosti, z ktorých chceme určiť najlepšiu voľbu a stĺpce matice budú pozostávať z rôznych kritérií, na základe ktorých sa budeme chcieť rozhodovať. Označovať ju budeme $A_{m\times n}$, pričom jej prvok a_{ij} bude predstavovať j-te kritérium i-tej možnosti.

ii. Stĺpcové normalizovanie matice A

$$\bar{a}_{ij} = \frac{a_{ij}}{\sqrt{\sum_{k=1}^{m}(a_{kj})^2}}$$

Čím zabezpečíme, že všetky hodnoty matice A budú v absolútnej hodnote z intervalu $\langle 0,1\rangle$.

iii. Výpočet váženej normalizovanej matice rozhodovania

Každému kritériu priradíme váhu tak, aby súčet všetkých váh bol rovný 1. Následne prenásobíme každý stĺpec prislúchajúci jednému kritériu príslušnou váhou.

$$\sum_{j=1}^{n} w_j = 1$$

$$x_{ij} = \bar{a}_{ij} * w_j$$

kde x_{ij} budú tvoriť prvky príslušnej matice, ktorú označíme X a $w_j \in \langle 0,1\rangle$.

iv. Určenie najlepšej a najhoršej potenciálnej možnosti

$$x_j^b = \max_{i=1,\dots,n} x_{ij}$$

$$x_j^w = \min_{i=1,\dots,n} x_{ij}$$

v. Výpočet euklidovskej vzdialeností pre všetky alternatívy od najlepšej a najhoršej potenciálnej možnosti

$$d_i^b = \sqrt{\sum_{j=1}^{n} \left(x_{ij} - x_j^b\right)^2}$$

$$d_i^w = \sqrt{\sum_{j=1}^{n} \left(x_{ij} - x_j^w\right)^2}$$

vi. Určenie podobnosti s najhoršou potenciálnou možnosťou

$$s_i = \frac{d_i^w}{d_i^w + d_i^b}$$

Prípad $s_i = 0$ nastáva iba, ak je vzdialenosť $d_i^w = 0$, čo znamená že daná alternatíva je zhodná s najhoršou potenciálnou možnosťou. Naopak ak je daná alternatíva zhodná s najlepšou potenciálnou možnosťou, potom je hodnota s_i rovná 1. Čím vyššia hodnota metriky s_i, tým je pre nás daná alternatíva lepšia

vii. Usporiadanie alternatív zostupne podľa hodnoty podobnosti
Ako posledné stačí dané možnosti usporiadať vzostupne podľa hodnoty s_i a zobrať ako prvú najlepšiu možnosť spomedzi všetkých alternatív.

Veľkou výhodou tohto algoritmu v našom prípade je to, že ak by jedna hrana mala veľmi nízku hodnotu jedného indexu, môže to veľmi dobre vykompenzovať veľmi dobrou hodnotou iného indexu. Jediné, čo nám ešte ostáva, je určiť váhy pre dané indexy.

Vždy je dobré určiť viacero možností na pridelenie váh indexom. My sme sa rozhodli aplikovať dve sady váh. Prvú budú tvoriť rovnomerné váhy. Čo v našom prípade ale nemusí byť vôbec ideálne. Ak sa vrátime do predchádzajúcej kapitoly a spomenieme si, ako boli jednotlivé indexy definované, tak ich vieme rozdeliť do troch skupín podľa druhu penalizácie. Indexy CN a PA tvoria skupinu indexov, ktoré spoločných susedov nepenalizujú vôbec. AA a RA zase tvoria skupinu, kde sú spoloční susedia penalizovaní podľa ich susedstiev. Zvyšné indexy penalizujú spoločných susedov podľa susedstiev dvojice vrcholov, pre ktoré chceme určiť podobnosť. Tieto tri skupiny nie sú ale vyrovnané podľa počtu indexov. Preto sa v takomto prípade môže veľmi ľahko stať, že najpočetnejšia skupina prehlasuje zvyšné dve. Z toho dôvodu

zavedieme druhú sadu váh, kde každej skupine priradíme určitú váhu, a zároveň rozložíme váhy v každej skupine medzi indexy rovnomerne. Keďže indexy CN a PA nie sú z pohľadu potenciálneho filtra najlepšie, pretože pridávajú hrany hlavne vrcholom s najväčším počtom už pridaných hrán, čo vyplýva už podľa toho ako sú tie dva indexy definované, rozhodli sme sa priradiť tejto skupine váhu rovnú 0.1. Zvyšným skupinám sme priradili rovnaké váhy rovné 0.45. Čím teda dostávame dve sady váh, tabuľka 1, ktoré budú vstupovať do TOPSIS algoritmu.

Tabuľka 1 Dve sady váh pre TOPSIS algoritmus

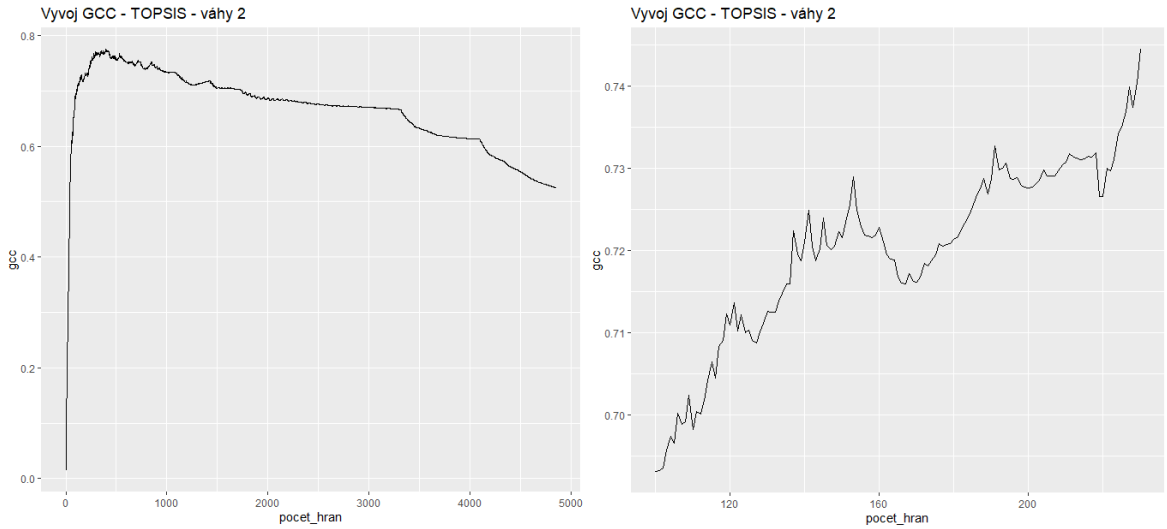
	CN	*PA*	*AA*	*RA*	*JC*	*SI*	*COS*	*LAS*	*HP*	*HD*	*LHN*
1	0.091	0.091	0.091	0.091	0.091	0.091	0.091	0.091	0.091	0.091	0.091
2	0.05	0.05	0.225	0.225	0.0643	0.0643	0.0643	0.0643	0.0643	0.0643	0.0643

Samotný algoritmus na filtrovanie informácie sa zmení iba v jednom kroku. Konkrétne v kroku iii. neskončíme pri obrátení hodnôt indexov pre jednotlivé hrany, ale na základe TOPSIS algoritmu a príslušnej sady váh určíme poradie hrán. V našom prípade bude najlepšie potenciálna možnosť reprezentovaná čisto iba maximami jednotlivých indexov. V kroku iv. potom pridáme hranu medzi vrcholy s najväčšou podobnosťou vzniku.

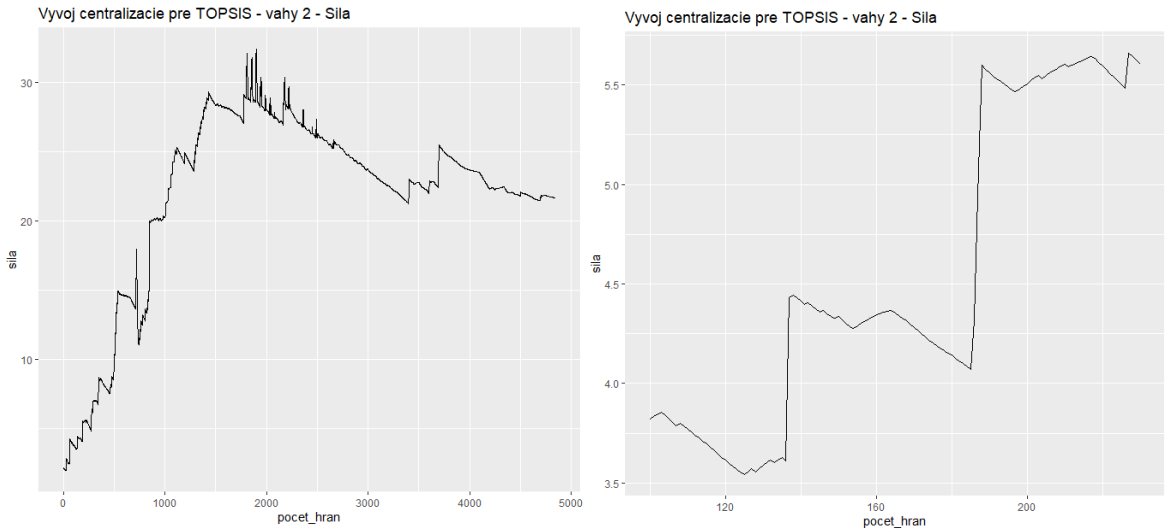
Pre druhú sadu váh si opäť demonštrujeme, ako určiť počet hrán. Rovnakým spôsobom, ako v prípade RA indexu, určíme nejaký veľmi dobrý bod kedy algoritmus zastaviť.

Rovnako sa najskôr pozrieme na vývoj globálneho koeficientu zhlukovania. Ten stúpa ešte strmšie ako v prípade indexu RA, a zároveň dosahuje aj vyššiu maximálnu hodnotu, obrázok 34. Opäť sa zameriame na to, aby počet hrán bol približne v rozmedzí, kde globálny koeficient zhlukovania ešte stúpa. Rozpätie teda vieme kvalitatívne určiť na $100 - 230$ pridaných hrán. Ďalej opätovne určíme veľmi dobrý bod podľa hodnôt centralizácie pre jednotlivé centrality, pričom zachováme myšlienku výberu tohto bodu. Z centralít sa javí ako veľmi dobrý bod 185, opätovne kvôli tomu, že tento bod určuje koniec klesania centralít z prvej skupiny a zároveň v ňom centrality z druhej skupiny nerastú (pre rozdelenie centralít do skupín pozri demonštrovanie filtru pre index RA v úvode tejto kapitoly), obrázky 35, 36, 37, 38, 39 a 40.

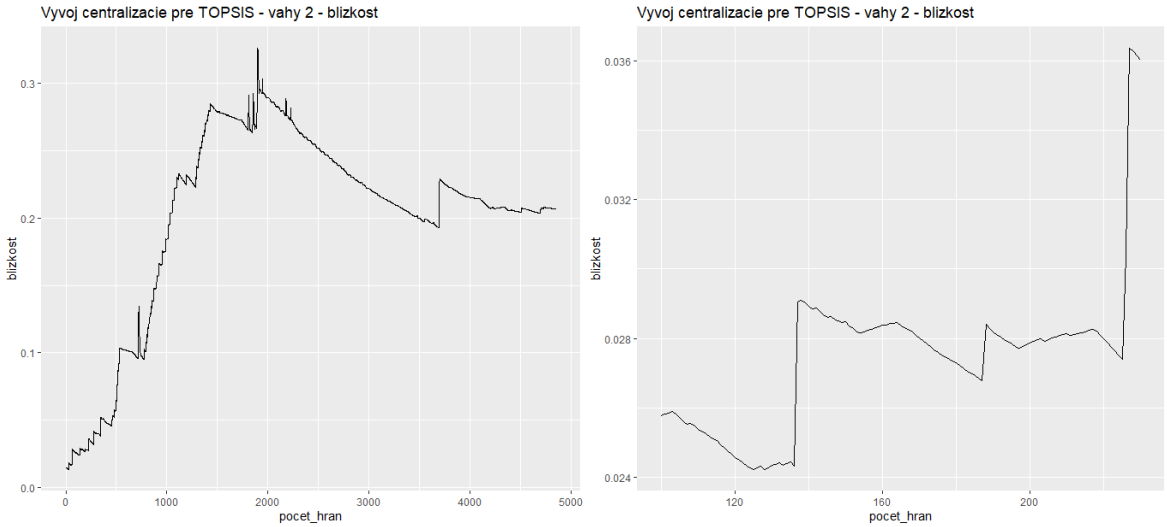
Výhodou takéhoto prístupu je, že využívame informáciu z viacerých pohľadov na podobnosť dvoch vrcholov v sieti. Jednou výzvou však ostáva zvolenie optimálneho rozloženia váh medzi indexy, poprípade nevyužívanie všetkých indexov naraz, ale iba nejakú časť z nich.



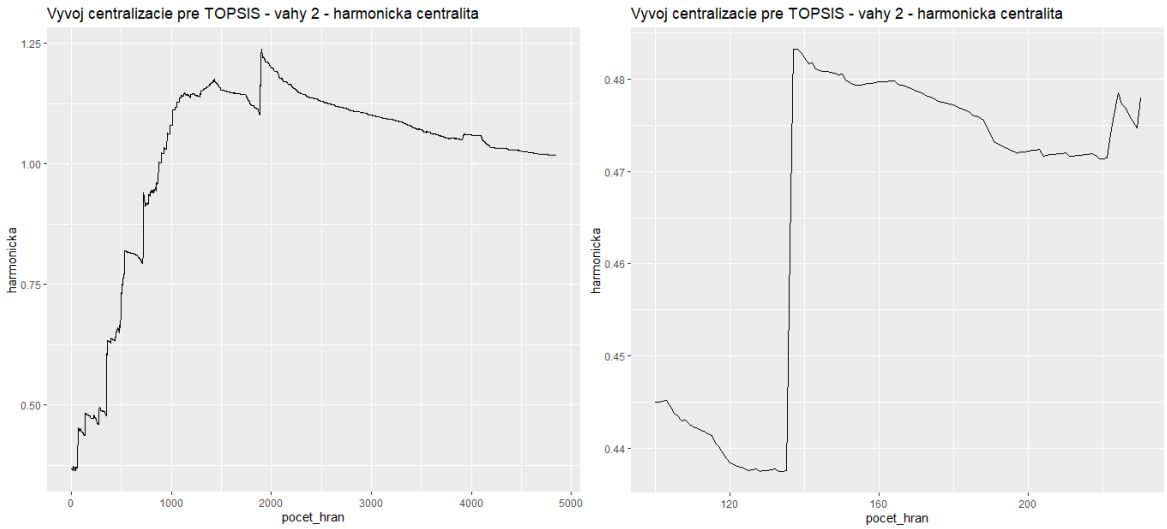
Obrázok 34 Vývoj hodnoty globálneho koeficientu zhlukovania pri pridávaní podľa TOPSIS2 – vpravo celkový priebeh, vľavo priblíženie na počet pridaných hrán v rozpätí 100-230



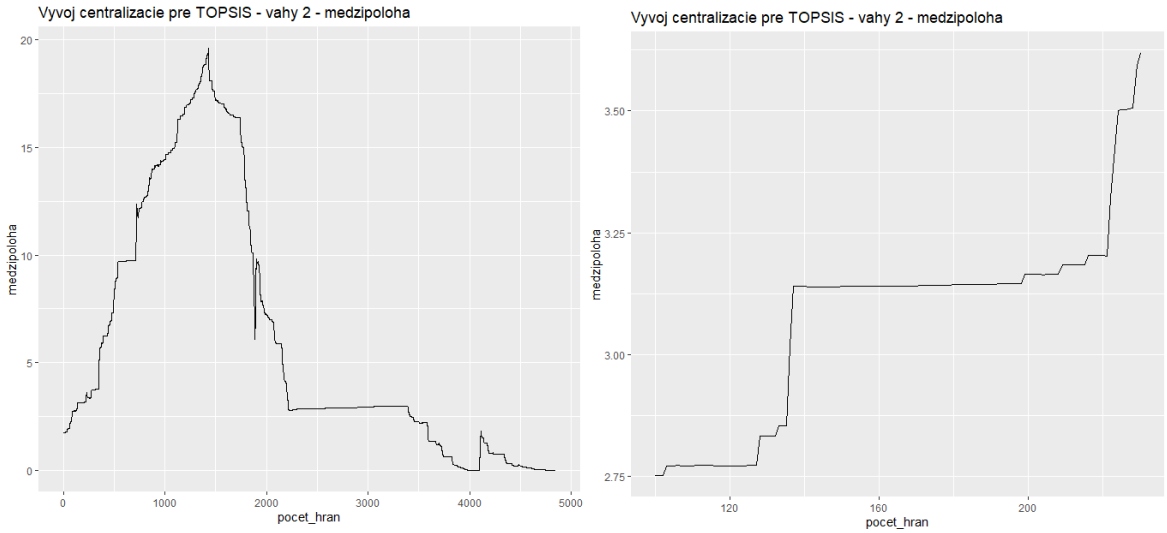
Obrázok 35 Vývoj hodnoty centralizácie sily pri pridávaní podľa TOPSIS2 – vpravo celkový priebeh, vľavo priblíženie na počet pridaných hrán v rozpätí 100-230



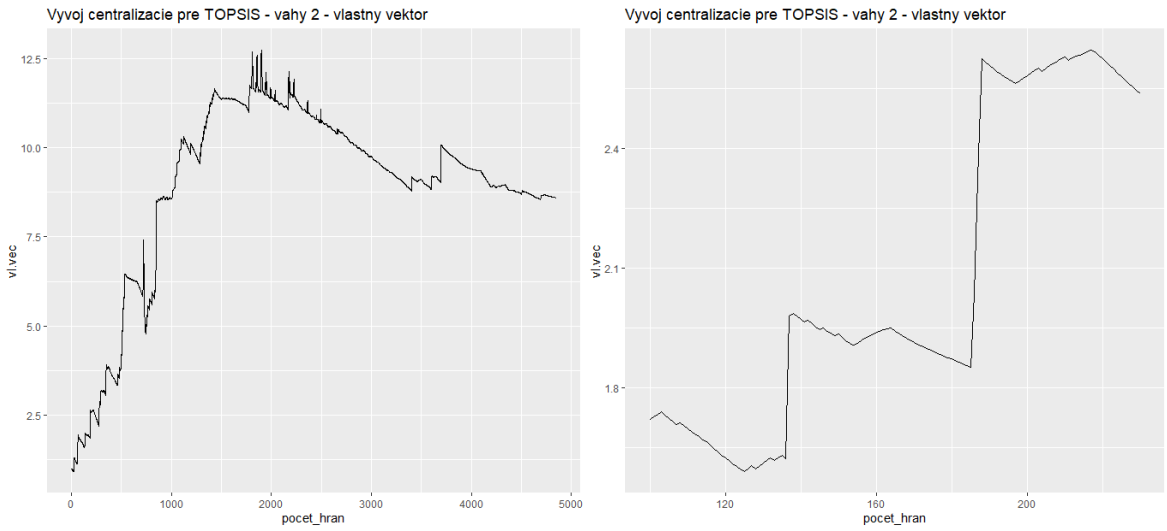
Obrázok 36 Vývoj hodnoty centralizácie blízkosti pri pridávaní podľa TOPSIS2 – vpravo celkový priebeh, vľavo priblíženie na počet pridaných hrán v rozpätí 100-230



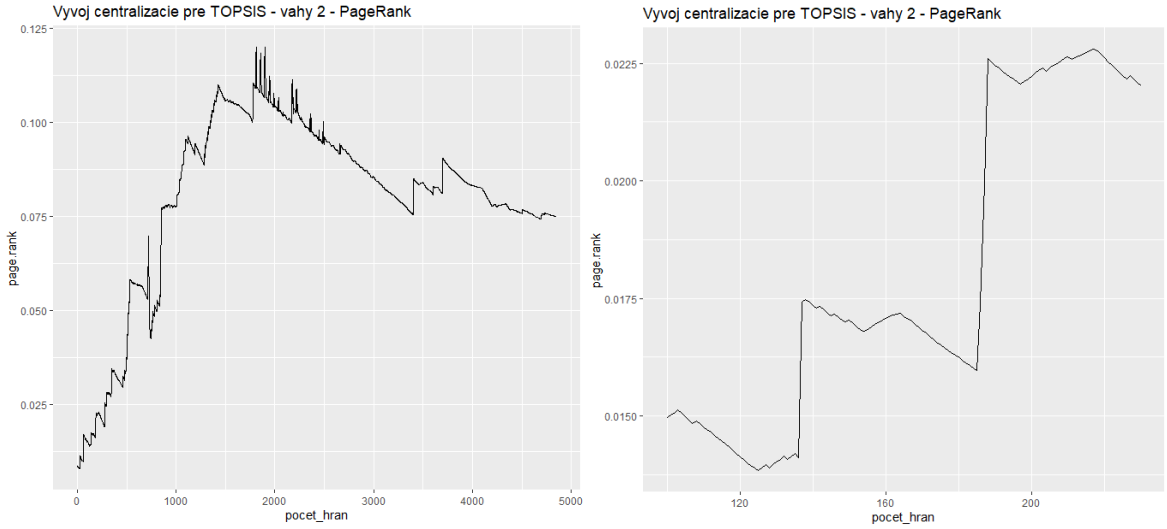
Obrázok 37 Vývoj hodnoty harmonickej centralizácie pri pridávaní podľa TOPSIS2 – vpravo celkový priebeh, vľavo priblíženie na počet pridaných hrán v rozpätí 100-230



Obrázok 38 Vývoj hodnoty centralizácie sily pri pridávaní podľa TOPSIS2 – vpravo celkový priebeh, vľavo priblíženie na počet pridaných hrán v rozpätí 100-230



Obrázok 39 Vývoj hodnoty centralizácie sily pri pridávaní podľa TOPSIS2 – vpravo celkový priebeh, vľavo priblíženie na počet pridaných hrán v rozpätí 100-230



Obrázok 40 Vývoj hodnoty centralizácie sily pri pridávaní podľa TOPSIS2 – vpravo celkový priebeh, vľavo priblíženie na počet pridaných hrán v rozpätí 100-230

Tabuľka 2 Počet hrán kvalitatívne zvolených pre každý prístup

Index	CN	AA	RA	PA	LHN	JC	SI
Počet hrán	355	229	218	276	215	184	188
Index	COS	HP	HD	LAS	TOPSIS1	TOPSIS2	TMFG
Počet hrán	211	202	203	206	226	225	294

4 Porovnanie prístupov

Pre porovnanie prístupov medzi sebou sme zvolili ako meradlo ich schopnosť zhlukovania. Všetky prístupy porovnáme medzi sebou, a to na základe toho ako veľmi podobné zhluky vytvárajú. Zároveň ich porovnáme aj so zhlukmi, ktoré vytvárajú filtre minimálnej kostry a TMFG. Chceme zistiť, či dodatočné pridávanie hrán nám dáva dodatočnú informáciu pre zhlukovanie vrcholov. To znamená, že ak pre nejaký prístup dostaneme rovnaké zhlukovanie ako v prípade minimálnej kostry, budeme daný prístup z pohľadu zhlukovania považovať za zbytočný, pretože nám nedal žiadnu dodatočnú informáciu. Druhým benchmarkom bude trojuholníkovo maximálne filtrovaný graf. Tento má oproti minimálnej kostre viac pridaných hrán. Opäť ani v tomto prípade nechceme dostať vysokú zhodu medzi zhlukovaním, ktoré vzniklo z metódy TMFG a zhlukovaním, ktoré vzniklo na základe niektorého nášho prístupu.

Na vytváranie zhlukov použijeme dva algoritmy, ktoré určujú komunity v sieťach. Konkrétne *Girvan-Newman* algoritmus a algoritmus *Walktrap*. Uvedieme si základnú myšlienku oboch algoritmov, ale keďže sa opäť nejedná o nosnú časť tejto práce, nebudeme sa im venovať podrobne.

4.1 Girvan-Newmanov algoritmus

Základná myšlienka algoritmu je veľmi jednoduchá a spočíva v nájdení a následnom odstránení hrán, ktoré jednotlivé komunity spájajú [64]. V praxi to znamená, že sa postupne z grafu odstraňujú dôležité hrany. Pre výpočet dôležitosti hrán je využité zovšeobecnenie centrality medzi polohy pre hrany v sieti. Táto centralita určuje v koľkých najkratších cestách sa daná hrana vyskytuje. Ak medzi vrcholmi existuje viacero najkratších ciest, potom každá najkratšia cesta dostane rovnakú váhu, tak aby súčet váh bol rovný 1. Ak sieť obsahuje komunity, ktoré sú prepojené iba pár hranami, potom veľký počet najkratších ciest bude viesť práve cez tieto hrany. Takže hrany, ktoré prepájajú komunity budú mať väčšiu hodnotu centrality. Odstránením týchto hrán oddelíme komunity medzi sebou a odhalíme komunitnú štruktúru siete. Algoritmus sa skladá z nasledovných štyroch krokov

i. Výpočet medzi polohy pre všetky hrany v sieti
ii. Odstránenie hrany s najväčšou hodnotou medzi polohy

iii. Opätovný výpočet medzi polohy pre všetky zvyšné hrany v sieti, ktorých centralita bola postihnutá predošlým odstránením hrany

Tu si treba rozmyslieť, čo to znamená. Ak sa nám po odstránení hrany rozpadne sieť na dve podsiete, je potreba stále počítať centralitu medzi polohy pre všetky zvyšné hrany. Avšak po ďalšom odstránení jednej z hrán, ostane jedna podsieť nedotknutá, a práve pre hrany v tejto podsieti nebudeme musieť opätovne počítať centralitu medzi polohy, keďže sa nezmení.

iv. Opakujeme krok ii. a iii. pokiaľ žiadna hrana neostane

Výsledkom daného algoritmu je dendogram, ktorý je vytvorený zhora dole. To znamená, že nazačiatku máme jeden veľký zhluk, ktorý postupne rozbíjame na menšie zhluky, až sa napokon dostaneme k zhlukom, ktoré tvoria iba jednotlivé vrcholy. Vetvy dendogramu predstavujú jednotlivé vrcholy grafu.

4.2 Algoritmus Walktrap

Walktrap algoritmus je založený na náhodnej prechádzke po sieti. Pod pojmom náhodná prechádzka môžeme chápať ľubovoľný pohyb po sieti, kde sa premiestňujeme z jedného vrcholu do niektorého z jeho susediacich vrcholov náhodne [65]. Postupnosť navštívených vrcholov môžeme chápať ako Markovov reťazec a jednotlivé vrcholy nám budú predstavovať stavy. Myšlienka algoritmu je, že popri náhodnej prechádzke sieťou máme tendenciu zotrvať v hustejších častiach siete, teda v jej komunitách. Na rozdiel od Girvan-Newman algoritmu sa dendogram vytvára zdola nahor. Čo znamená, že začíname s tým, že každý zhluk obsahuje práve jeden vrchol. Postupne podľa kritéria založeného na vzdialenosti zhlukov, vyberieme jednu dvojicu zhlukov, ktorú spojíme do jedného zhluku. Tento postup zopakujeme, pokiaľ nevytvoríme iba jeden zhluk so všetkými vrcholmi. Dvojice komunít, ktoré sa následne spoja do jednej, vyberáme pomocou Wardovho algoritmu [66].

Na porovnanie výsledných zhlukov použijeme dve metriky. Jednu, ktorú sme si sami zostavili a upravený Randov index [67]. Keďže všetky prístupy, či už podľa jednotlivých indexov alebo s využitím TOPSIS algoritmu, vychádzajú z minimálnej kostry, z istej časti je zhlukovanie ovplyvnené práve minimálnou kostrou. Preto sa naša metrika bude opierať iba o porovnanie dvojíc, ktoré skončili pri daných prístupoch spolu v zhluku, keďže počet dvojíc, ktoré neskončili spolu v zhluku by bolo príliš veľa

a značne by to porovnanie vychýlilo. Metrika je jednoduchá a zároveň veľmi intuitívna. Označme si Z maticu zhlukovania v zmysle

$$z_{ij} = \begin{cases} 1, & vrcholy\ v_i\ a\ v_j\ patria\ do\ rovnakého\ zhluku \\ 0, & vrcholy\ v_i\ a\ v_j\ nepatria\ do\ rovnakého\ zhluku \end{cases}$$

Potom vieme podobnosť zhlukovania vypočítať podľa vzťahu

$$S_{12} = \frac{\sum_{i,j=1}^{n} (Z_1 \otimes Z_2)_{ij}}{\sum_{i,j=1}^{n} (Z_1 + Z_2 - Z_1 \otimes Z_2)_{ij}}$$

pričom $Z_1 \otimes Z_2$ označujeme pozložkové násobenie matíc. Menovateľ predstavuje počet všetkých dvojíc, ktoré sa v jednom aj v druhom zhlukovaní dostali spolu do zhluku. Čitateľ predstavuje celkový počet dvojíc v jednom a druhom zhlukovaní, pričom každá dvojica je zarátaná práve raz. Podobnosť $S_{12} \in \langle 0,1 \rangle$, pričom $S_{12} = 1$ práve vtedy, keď sú obe zhlukovania totožné a $S_{12} = 0$ práve vtedy, keď nemajú ani jednu spoločnú dvojicu vrcholov.

4.3 Upravený Randov index

Označme dve zhlukovania $X = \{X_1, X_2, \dots, X_r\}$ a $Y = \{Y_1, Y_2, \dots, Y_s\}$, kde prvky X_i, respektíve Y_i predstavujú množiny tých vrcholov, ktoré patria do zhluku i. Ďalej si označme $n_{ij} = |X_i \cap Y_j|$. Prienik X a Y vieme zosumarizovať podľa nasledovnej kontingenčnej tabuľky

$X \setminus Y$	Y_1	Y_2	…	Y_s	súčet
X_1	n_{11}	n_{12}	…	n_{1s}	a_1
X_2	n_{21}	n_{22}	…	n_{2s}	a_2
…	…	…	…	…	…
X_r	n_{r1}	n_{r2}	…	n_{rs}	a_r
súčet	b_1	b_2	…	b_s	

Potom vieme upravený Randov index definovať ako [67]

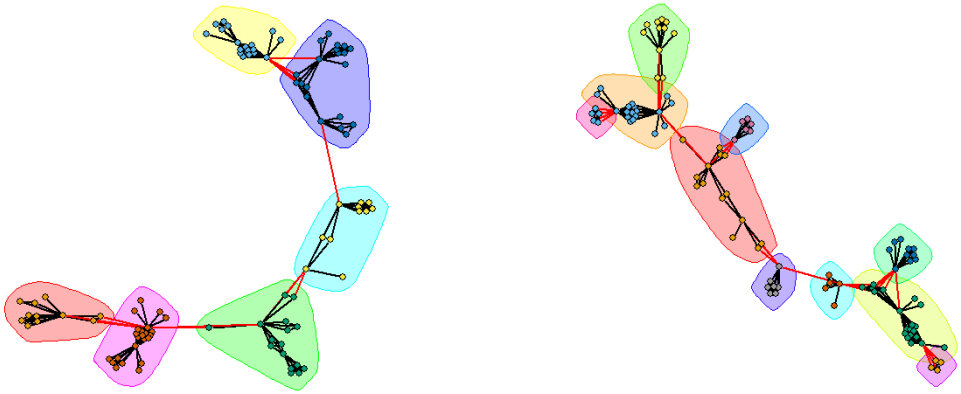
$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}}$$

pričom n predstavuje počet vrcholov.

Zhlukovania budeme porovnávať najskôr pre siete s počtom hrán, ktoré sme určili podľa globálneho koeficientu zhlukovania a centralizácie, následne pre siete s fixným počtom hrán, ktorý bude totožný s počtom hrán, ktoré nám vráti TMFG, čiže $3n - 6$. V druhom prípade sa pozrieme dodatočne aj na priemernú váženú hustotu zhlukov, ktoré jednotlivé prístupy vytvoria. Tá nám dá informáciu, okrem toho, ako husté zhluky jednotlivé prístupy vytvárajú, aj o podobností jednotlivých prístupov. Počítať ju budeme podľa nasledovného vzťahu

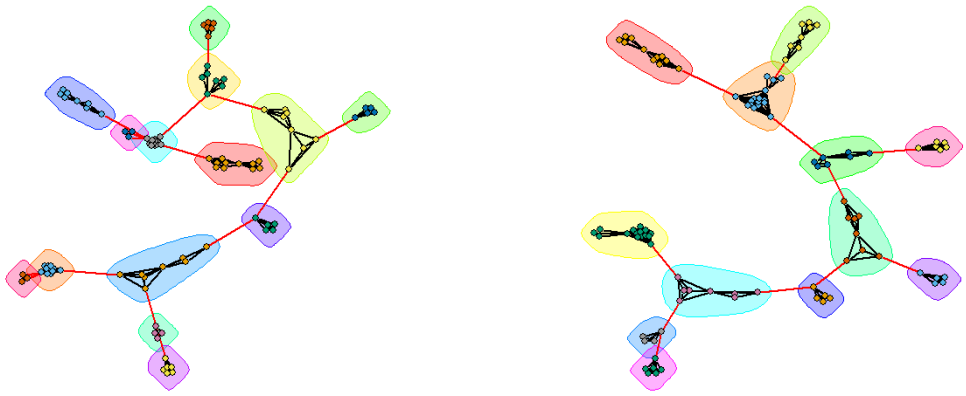
$$D = \sum_i w_i D_i$$

kde D_i predstavuje hustotu i-tého zhluku a $w_i = \frac{Počet\ vrcholov\ v\ zhluku\ i}{Počet\ všetkých\ vrcholov\ v\ sieti}$. Ľahko sa dá vidieť, že $\sum_i w_i = 1$.



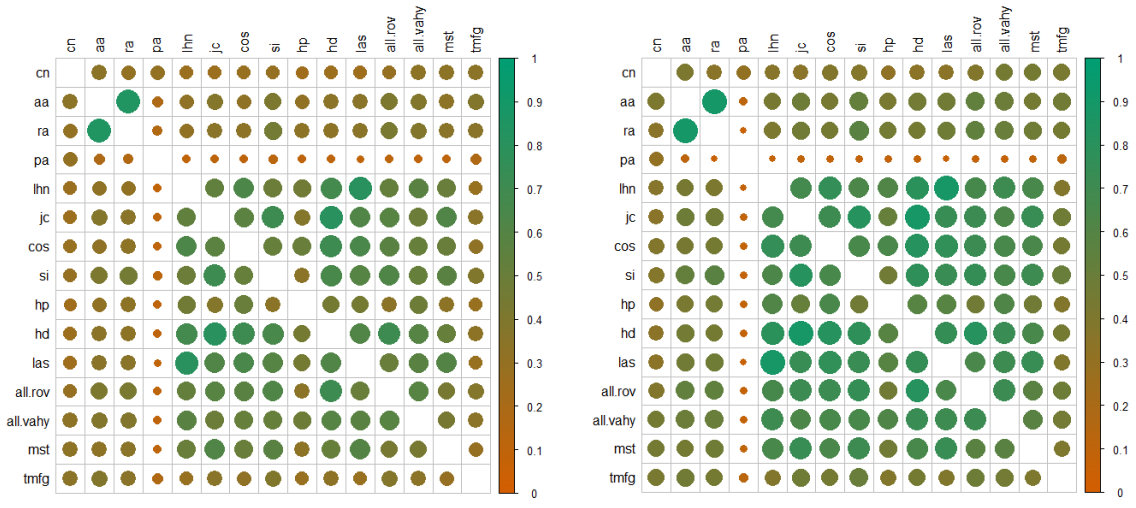
Obrázok 41 Zhlukovanie podľa RA – Girvan-Newmanov algoritmus vľavo, Walktrap algoritmus vpravo

Začneme porovnávaním pri nami určenom počte hrán podľa tabuľky 2. Ak sa pozrieme na zhlukovania vytvorené podľa Girvan-Newmanovho algoritmu, môžeme si všimnúť, že indexy AA a RA tvoria veľmi podobné zhlukovania. Indexy CN a PA sú naopak málo podobné všetkým ostatným prístupom, čo je očakávané, vzhľadom na ich definície. Indexy JC, SI, LHN, HP, HD, LAS, COS a oba prístupy využívajúce TOPSIS

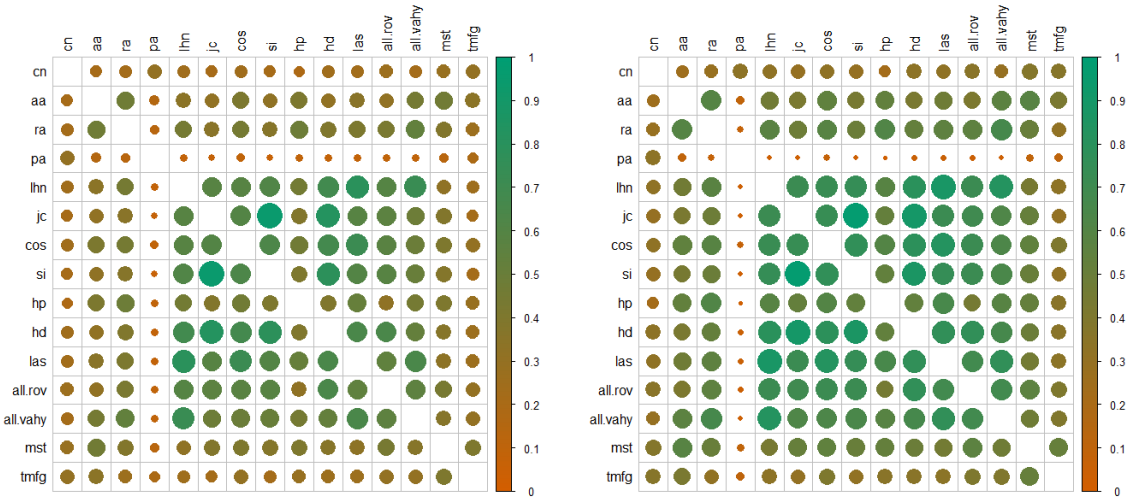


Obrázok 42 zhlukovanie podľa TOPSIS 2 - Girvan-Newmanov algoritmus vľavo, Walktrap algoritmus vpravo

algoritmus taktiež tvoria veľmi podobné zhlukovania. Čo je ale dôležité si všimnúť je, že táto skupinka prístupov tvorí podobné zhlukovania ako minimálna kostra. Čo znamená, že vzhľadom na zhlukovanie nám tieto prístupy nedávajú, respektíve nefiltrujú oveľa väčšiu informáciu oproti minimálnej kostre. Naopak indexy RA a PA tvoria výrazné odlišné zhluky na rozdiel od minimálnej kostry, čo znamená, že získavame informáciu navyše, a zároveň sú rozdielne aj oproti zhlukom tvorených TMFG, to vieme rovnako interpretovať, ako získavanie novej informácie oproti filtrom minimálna kostra a TMFG. Ak sa pozrieme na zhluky vytvorené algoritmom Walktrap, môžeme pozorovať veľmi podobné výsledky, až na jeden rozdiel, ktorým je rozbitie zhlukovania RA do väčšieho počtu zhlukov, ako tomu bolo pri Girvan-Newmanovmu algoritme. Opäť ale môžeme pozorovať nízku podobnosť so zhlukovaním minimálnej kostry a TMFG, obrázky 43 a 44.

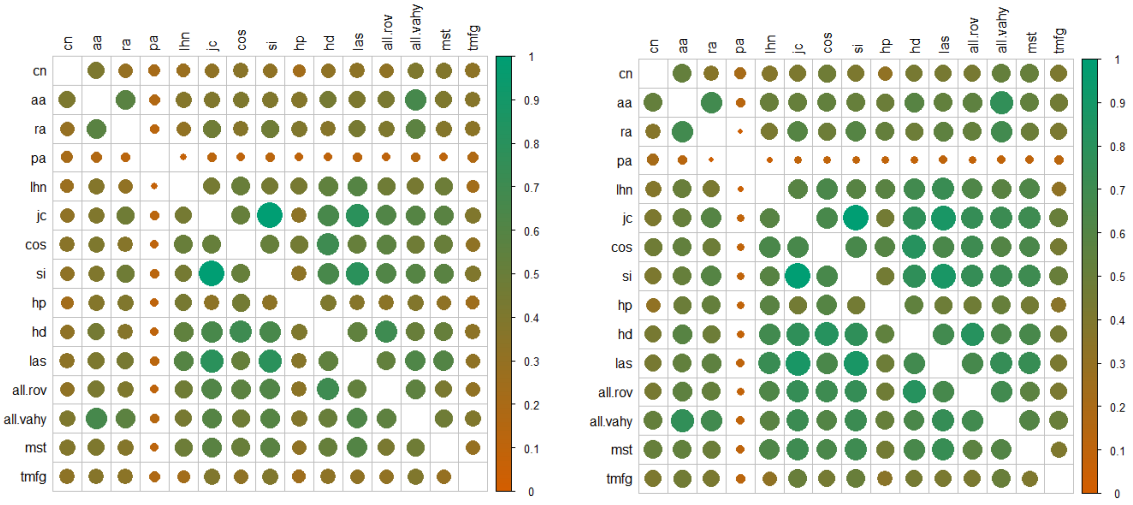


Obrázok 43 – Podobnosť zhlukovania – Girvan_Newmanov algoritmus – nami zvolený počet hrán – vľavo náš index podobnosti, vpravo upravený Randov index

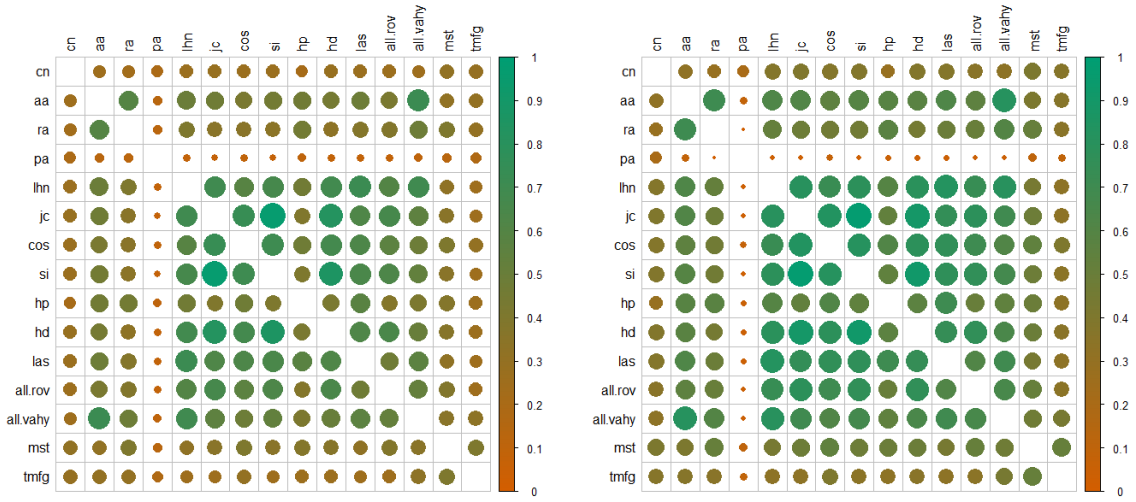


Obrázok 44 – Podobnosť zhlukovania – Walktrap algoritmus – nami zvolený počet hrán – vľavo náš index podobnosti, vpravo upravený Randov index

Pri zafixovanom počte hrán na hodnotu 294, môžeme pozorovať skoro totožné výsledky, s jedným veľkým rozdielom, a tým je veľká podobnosť medzi zhlukovaním podľa indexu AA a TOPSIS2 prístupu. Opäť môžeme pozorovať menšiu podobnosť, ako so zhlukovaním minimálnej kostry, tak aj zhlukovaním TMFG. Teda pomocou nami vytvorených prístupov dostávame väčšiu informáciu oproti minimálnej kostre, a zároveň je táto informácia iný, ako tá, ktorú dostávame pomocou TMFG, obrázky 45 a 46.



Obrázok 45 – Podobnosť zhlukovania – Girvan_Newmanov algoritmus – fixný počet hrán (294) – vľavo náš index podobnosti, vpravo upravený Randov index



Obrázok 46 – Podobnosť zhlukovania – Walktrap algoritmus – fixný počet hrán (294) – vľavo náš index podobnosti, vpravo upravený Randov index

Ako posledné sa pozrieme na váženú priemernú hustotu zhlukov pri fixovanom počte pridaných hrán, tabuľky 3 a 4.

Tabuľka 3 Vážená priemerná hustota zhlukov pri fixnom počte hrán – Girvan-Newmanov algoritmus

INDEX	CN	AA	RA	PA	LHN	JC	COS
Priemerná hustota	0.366	0.442	0.372	0.414	1	0.618	0.912
INDEX	**SI**	**HP**	**HD**	**LAS**	**TOPSIS1**	**TOPSIS2**	**TMFG**
Priemerná hustota	0.618	0.819	0.816	0.668	0.853	0.519	0.367

Tabuľka 4 Vážená priemerná hustota zhlukov pri fixnom počte hrán – Walktrap algoritmus

INDEX	CN	AA	RA	PA	LHN	JC	COS
Priemerná hustota	0.487	0.502	0.374	0.534	0.756	0.872	0.912
INDEX	**SI**	**HP**	**HD**	**LAS**	**TOPSIS1**	**TOPSIS2**	**TMFG**
Priemerná hustota	0.909	0.713	0.845	0.820	0.853	0.613	0.432

Čo z toho vieme vydedukovať je to, že indexy CN, PA, AA, RA a prístup TOPSIS2 produkujú väčšie a menej husté zhluky, podobne ako TMFG. Naopak zvyšné indexy a prístup TOPSIS1 produkujú viacero menších zhlukov, ktoré sú pomerné husté.

Berúc do úvahy všetky porovnávania, môžeme prehlásiť, že vzhľadom na zhlukovanie siete, sa indexy AA, RA a prístup TOPSIS2 sa javia ako lepšia voľba pri vytváraní filtrovanej siete oproti zvyšným indexom a prístupu TOPSIS1. Vychádzame z toho, že poskytujú pomerne rozšírenie informácie oproti minimálnej kostre, čo je veľmi vhodné, keďže minimálna kostra a informácia s ňou spätá je obsiahnutá vo všetkých filtrovaných sieťach, keďže je minimálna kostra štartovacou podsieťou algoritmov. Zároveň poskytujú rozdielnu informáciu oproti TMFG, pričom vytvárajú podobne veľké a husté zhluky. Problémy zvyšných prístupov sú, že oproti minimálnej kostre siete nedostávame oveľa väčšiu informáciu, a zároveň vytvárajú viacero menších a hustejších zhlukov, čo by taktiež mohlo znamenať, že nepoodokrývajú hlavnú štruktúru siete.

Záver

Cieľom diplomovej práce bola predikcia nepozorovaných alebo budúcich hrán v znamienkových vážených sieťach. Opierali sme sa hlavne o existujúcu literatúru, ktorá sa zaoberá problémom predikcie hrán v binárnych, poprípade vo vážených sieťach.

Diplomová práca je členená do štyroch kapitol.

V prvej kapitole sme sa venovali dôležitým veciam z teórie sietí. Ako prvé sme si zadefinovali, čo to sieť vôbec je. Následne sme si prešli prvotné rozdelenie sietí v závislosti na povahe vzťahov medzi vrcholmi siete. Uviedli sme si pojem finančnej siete, konkrétne korelačnej siete vytvorenej pomocou korelácií výnosov akcií. Vzhľadom na daný typ siete, ktorým je plná sieť, sme si priblížili aj najpoužívanejšie algoritmy na filtrovanie informácie zo sieti. Na záver kapitoly sme si zaviedli užitočné definície z teórie sietí, konkrétne koeficient zhlukovania a centralitu vrcholov.

Druhá kapitola bola venovaná predikcií hrán. Začali sme krátkym prehľadom už existujúcich prístupov a rozdelili sme ich podľa princípu na akom sú založené. Ďalej sme sa venovali už iba jednej skupine prístupov. Tvoria ju prístupy využívajúce lokálnu topológiu siete. Hlavnou myšlienkou týchto prístupov bolo odhadovanie podobnosti medzi vrcholmi s tým, že čím viac sú dva vrcholy v istom zmysle podobné, tým je väčšia pravdepodobnosť, že v budúcnosti budú spojené hranou. Tieto metódy boli poväčšine využívané na predikciu hrán v binárnych sieťach. Vznikli aj zovšeobecnenia pre vážené siete, nie však pre všetky prístupy z tejto kategórie. Práve prístupy využívajúce lokálnu topológiu siete poslúžili ako základ pre vytvorenie metód na predikciu hrán v znamienkových vážených sieťach. Vybrali sme najpoužívanejšie indexy, ktoré sa v literatúre objavovali, a zovšeobecnili sme ich tak, aby ich bolo možné použiť aj na prípad znamienkových sietí. Pri zovšeobecňovaní sme brali do úvahy viaceré faktory. Najdôležitejším bolo, aby sa hlavná myšlienka indexov podobnosti zachovala. To taktiež znamená, že ak by sa zovšeobecnené indexy nepoužili na znamienkové siete, ale iba na binárne alebo vážené siete, výsledný zovšeobecnený index by mal byť veľmi dobrou aproximáciou pôvodného indexu. Ďalším dôležitým faktorom bolo znamienko váhy. Chceli sme docieliť, aby pomocou zovšeobecnených indexov bolo možné merať aj akúsi anti-podobnosť, čo predstavuje podobnosť so záporným znamienkom. Okrem iného sme kládli dôraz aj na spojitosť daných indexov, a to z dôvodu, aby mala zmena

v hodnote váhy nejakej hrany nespôsobila veľkú zmenu v hodnote indexu. Na záver kapitoly sme analyzovali správanie sa indexov medzi sebou, a rovnako aj ich synergiu s pôvodnými hodnotami korelácií. Z príkladov filtrovaných sietí a plnej siete sa zdá, že zovšeobecnené indexy vierohodne odzrkadľujú podobnosť vrcholov.

V tretej kapitole sme sa použili na jedno z potenciálnych využití predikcie hrán v korelačných sieťach. Využívajúc zovšeobecnené indexy podobnosti sme vytvorili jednoduchý filter, ktorý by nám mohol pomôcť k odokrytiu dôležitej informácie, ktorá v prípade plnej sieti nemusí byť dostatočne viditeľná. Najskôr sme využívali každý index podobnosti samostatne, no neskôr sme sa rozhodli využiť informáciu zo všetkých dostupných indexov naraz. Využili sme pritom algoritmus TOPSIS, ktorý slúži na rozhodovanie sa podľa viacerých kritérií. Odvodili sme dve sady váh. Prvú sadu tvorili rovnomerné váhy, kde každý index mal rovnaké slovo pri rozhodovaní o najpodobnejšej dvojici vrcholov. Druhú sadu váh sme odvodili na základe podobností indexov vzhľadom na ich definíciu, konkrétne na druh penalizácie, ktorý využívali. Pre všetky prípady filtrovania sme určili veľmi dobrý bod na ukončenie algoritmu podľa vývoja globálneho koeficientu zhlukovania a vývoja centralizácií.

V poslednej kapitole sme sa venovali vyhodnoteniu jednotlivých indexov, ako aj prístupov využívajúcich TOPSIS algoritmus, vzhľadom na zhlukovanie vrcholov. Použili sme dva algoritmy, ktoré slúžia na vyhľadávanie komunít v sieťach, Girvan-Newmanov algoritmus a algoritmus Walktrap. Taktiež sme si zaviedli metriku, ktorá využíva dvojice vrcholov, ktoré sa nachádzajú spolu v zhluku, na porovnanie výsledných zhlukovaní jednotlivých prístupov. Najskôr sme porovnali zhlukovania pre prístupy s rôznymi počtami pridaných hrán. Neskôr sme sa pozreli aj na vzniknuté zhlukovania, ak bol počet pridaných hrán fixný. Doplnkovo sme sa v tomto prípade pozreli aj na priemernú hustotu vzniknutých zhlukovaní. Určili sme, ktoré prístupy sa zdajú byť lepšie, respektíve horšie, vzhľadom na pridanú informáciu pri zhlukovaní oproti minimálnej kostre grafu a TMFG.

Do budúcna vidíme hneď niekoľko možností, ako v práci pokračovať. Jednou z nich je zamerať sa na zovšeobecnenie aj iných prístupov, ako tých využívajúcich lokálnu topológiu sietí, pre predikciu hrán v znamienkových vážených sieťach. Rovnako užitočné by mohlo byť otestovať funkčnosť nami zovšeobecnených indexov na iný typ sietí, keďže korelačné siete sú typické tým, že hodnoty váh sú v absolútnej

hodnote menšie ako 1. Zaujímavou aplikáciou nami zovšeobecnených indexov by mohlo byť namiesto predikcie hrán sledovanie, ako by sa mala zmeniť štruktúra siete, ak nejakú hranu prudko zmeníme, v zmysle kladnú hranu zmeníme na zápornú. To by následne mohlo nájsť svoje uplatnenie v sociálnych sieťach, pri zmene vzťahov medzi dvomi osobami a vplyv tejto zmeny na zvyšné vzťahy v sieti. Rovnako zaujímavé by mohlo byť vylepšenie filtra. Tú sa núka hneď niekoľko možností. Napríklad namiesto matice vzdialeností, respektíve matice korelácií, ktoré vstupujú do algoritmov, použiť pri výpočte minimálnej kostry alebo TMFG maticu, ktorej prvky budú tvoriť podobnosť vrcholov určenej z pôvodnej siete. Priamym vylepšením nami predloženého filtra by zasa mohlo byť využitie aj pôvodnej siete pri pridávaní hrán. Tým máme namysli, že namiesto brania do úvahy iba podobnosť vrcholov na základe vzniknutých podsietí, by sme brali do úvahy aj podobnosť vrcholov, vypočítaných z pôvodnej siete, kde by buď boli všetky pôvodné hrany alebo iba tie, ktoré sa ešte do filtrovanej siete nedostali.

Zoznam použitej literatúry

[1] Li, Xin, and Hsinchun Chen, Recommendation as link prediction in bipartite graphs: A graph kernel-based machine learning approach, Decision Support Systems 54.2, 2013, p 880-890

https://www.sciencedirect.com/science/article/pii/S0167923612002540

[2] Barbieri, Nicola, Francesco Bonchi, and Giuseppe Manco, Who to follow and why: link prediction with explanations, Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014

https://dl.acm.org/doi/pdf/10.1145/2623330.2623733

[3] Su, Zhan, et al., Link prediction in recommender systems based on vector similarity, Physica A: Statistical Mechanics and its Applications 560, 2020, 125154

https://www.sciencedirect.com/science/article/pii/S0378437120306038

[4] Wang, Liang, Ke Hu, and Yi Tang, Robustness of link-prediction algorithm based on similarity and application to biological networks, Current Bioinformatics 9.3, 2014, p 246-252

https://arxiv.org/ftp/arxiv/papers/1302/1302.5878.pdf

[5] Barzel, Baruch, and Albert-László Barabási, Network link prediction by global silencing of indirect correlations, Nature biotechnology 31.8, 2013, p 720-725

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3740009/

[6] Sulaimany, Sadegh, Mohammad Khansari, and Ali Masoudi-Nejad, Link prediction potentials for biological networks, International Journal of Data Mining and Bioinformatics 20.2, 2018, p 161-184

https://www.inderscience.com/info/inarticle.php?artid=93684

[7] Term H., Reinert G., Statistical Interence for Networks. Boston: 2009. http://www.stats.ox.ac.uk/wpcontent/uploads/2018/08/Statistical_Inference_for_Networks.pdf.

[8] Mantegna, Rosario N., Hierarchical structure in financial markets, The European Physical Journal B-Condensed Matter and Complex Systems 11.1, 1999, p 193-197

https://link.springer.com/content/pdf/10.1007/s100510050929.pdf

[9] Onnela, J-P., Kimmo Kaski, and Janos Kertész, Clustering and information in correlation based financial networks, The European Physical Journal B 38.2, 2004, p 353-362

https://link.springer.com/content/pdf/10.1140/epjb/e2004-00128-7.pdf

[10] Tang, Yong, et al., How do the global stock markets Influence one another? Evidence from finance big data and granger causality directed network, International Journal of Electronic Commerce 23.1, 2019, p 85-109

https://doc.rero.ch/record/324144/files/zha_hdg.pdf

[11] Sensoy, Ahmet, and Benjamin M. Tabak, Dynamic spanning trees in stock market networks: The case of Asia-Pacific, Physica A: Statistical Mechanics and its Applications 414, 2014, p 387-402

https://www.sciencedirect.com/science/article/pii/S0378437114006517

[12] Namaki, A., et al., Network analysis of a financial market based on genuine correlation and threshold method, Physica A: Statistical Mechanics and its Applications 390.21-22, 2011, p 3835-3841

https://www.sciencedirect.com/science/article/pii/S0378437111004808

[13] Nobi, Ashadun, et al., Effects of global financial crisis on network structure in a local stock market, Physica A: Statistical Mechanics and its Applications 407, 2014, p 135-143

https://www.sciencedirect.com/science/article/pii/S0378437114002945

[14] Isogai, Takashi., Building a dynamic correlation network for fat-tailed financial asset returns, Applied network science 1.1, 2016, p 1-24

https://appliednetsci.springeropen.com/articles/10.1007/s41109-016-0008-x

[15] Boginski, Vladimir, Sergiy Butenko, and Panos M. Pardalos, On structural properties of the market graph, Innovations in financial and economic networks 48, 2003, p 29-35.

https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.59.3803&rep=rep1&type=pdf

[16] Graham, Ronald L., and Pavol Hell. "On the history of the minimum spanning tree problem." Annals of the History of Computing 7.1,1985, p 43-57

https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4392963&tag=1

[17] Miccichè, Salvatore, et al., Degree stability of a minimum spanning tree of price return and volatility., Physica A: Statistical Mechanics and its Applications 324.1-2, 2003, p 66-73.

https://www.sciencedirect.com/science/article/pii/S0378437103000025

[18] Wang, Gang-Jin, et al., Similarity measure and topology evolution of foreign exchange markets using dynamic time warping method: Evidence from minimal spanning tree, Physica A: Statistical Mechanics and its Applications 391.16, 2012, p 4136-4146

https://www.sciencedirect.com/science/article/pii/S0378437112002889

[19] Tumminello, Michele, et al., A tool for filtering information in complex systems, Proceedings of the National Academy of Sciences 102.30, 2005, p 10421-10426

https://arxiv.org/pdf/cond-mat/0501335.pdf

[20] Tumminello, Michele, et al., Correlation based networks of equity returns sampled at different time horizons, The European Physical Journal B 55.2, 2007, p 209-217

https://link.springer.com/content/pdf/10.1140/epjb/e2006-00414-4.pdf

[21] Birch, Jenna, Athanasios A. Pantelous, and Konstantin Zuev., The maximum number of 3-and 4-cliques within a planar maximally filtered graph, Physica A: Statistical Mechanics and its Applications 417, 2015, p 221-229

https://www.sciencedirect.com/science/article/pii/S0378437114007699#bbr000055

[22] Massara, Guido Previde, Tiziana Di Matteo, and Tomaso Aste., Network filtering for big data: Triangulated maximally filtered graph, Journal of complex Networks 5.2, 2017, p 161-178

https://academic.oup.com/comnet/article/5/2/161/2555365

[23] Watts, Duncan J., and Steven H. Strogatz., Collective dynamics of 'small-world'networks, nature 393.6684, 1998, p 440-442

https://edisciplinas.usp.br/pluginfile.php/4205021/mod_resource/content/1/NAture_Watts_Strogatz.pdf

[24] Costantini, Giulio, and Marco Perugini., Generalization of clustering coefficients to signed correlation networks, PloS one 9.2, 2014

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0088669

[25] Rodrigues, Francisco Aparecido., Network centrality: an introduction, A mathematical modeling approach from nonlinear dynamics to complex systems. Springer, 2019, p 177-196

https://link.springer.com/chapter/10.1007/978-3-319-78512-7_10

[26] Deverashetti, Mounika, and Samit Kumar Pradhan., Identification of topologies by using harmonic centrality in huge social networks, 2018 3rd International Conference on Communication and Electronics Systems (ICCES). IEEE, 2018

https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8724042

[27] Krnc, Matjaž, and Riste Škrekovski., Group centralization of network indices, Discrete Applied Mathematics 186, 2015, p 147-157

https://www.sciencedirect.com/science/article/pii/S0166218X15000086

[28] Katz, Leo., A new status index derived from sociometric analysis, Psychometrika 18.1, 1953, p 39-43

https://link.springer.com/content/pdf/10.1007/BF02289026.pdf

[29] Leicht, Elizabeth A., Petter Holme, and Mark EJ Newman., Vertex similarity in networks, Physical Review E 73.2, 2006, 026120

https://arxiv.org/pdf/physics/0510143.pdf

[30] Jeh, Glen, and Jennifer Widom., Simrank: a measure of structural-context similarity, Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 2002

https://dl.acm.org/doi/pdf/10.1145/775047.775126

[31] Fouss, Francois, et al., Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation, IEEE Transactions on knowledge and data engineering 19.3, 2007, p 355-369

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4072747

[32] Lü, Linyuan, Ci-Hang Jin, and Tao Zhou., Similarity index based on local paths for link prediction of complex network, Physical Review E 80.4, 2009, 046122

http://image.sciencenet.cn/olddata/kexue.com.cn/upload/blog/file/2009/10/2009102824850329588.pdf

[33] Liu, Weiping, and Linyuan Lü., Link prediction based on local random walk, EPL (Europhysics Letters) 89.5, 2010, 58007

https://arxiv.org/pdf/1001.2467.pdf

[34] Papadimitriou, Alexis, Panagiotis Symeonidis, and Yannis Manolopoulos., Fast and accurate link prediction in social networking systems, Journal of Systems and Software 85.9, 2012, p 2119-2132

https://www.sciencedirect.com/science/article/pii/S0164121212001069

[35] Clauset, Aaron, Cristopher Moore, and Mark EJ Newman., Hierarchical structure and the prediction of missing links in networks, Nature 453.7191, 2008, p 98-101

https://arxiv.org/pdf/0811.0484.pdf

[36] Peixoto, Tiago P., Hierarchical block structures and high-resolution model selection in large networks, Physical Review X 4.1, 2014, 011047

https://journals.aps.org/prx/pdf/10.1103/PhysRevX.4.011047

[37] Huang, Zan., Link prediction based on graph topology: The predictive value of generalized clustering coefficient, Available at SSRN 1634014, 2010

https://ailab.ijs.si/dunja/LinkKDD2006/Papers/huang.pdf

[38] Jeffrey A. Ryan and Joshua M. Ulrich (2020). quantmod: Quantitative Financial Modelling Framework. R package version 0.4.17.

https://CRAN.R-project.org/package=quantmod

[39] Duan, Liang, et al., An ensemble approach to link prediction, IEEE Transactions on Knowledge and Data Engineering 29.11, 2017, p 2402-2416

https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7987705

[40] Ahmed, Amr, et al., Distributed large-scale natural graph factorization, Proceedings of the 22nd international conference on World Wide Web, 2013

https://dl.acm.org/doi/pdf/10.1145/2488388.2488393

[41] Cao, Shaosheng, Wei Lu, and Qiongkai Xu., Grarep: Learning graph representations with global structural information, Proceedings of the 24th ACM international on conference on information and knowledge management, 2015

https://dl.acm.org/doi/pdf/10.1145/2806416.2806512

[42] Keikha, Mohammad Mehdi, Maseud Rahgozar, and Masoud Asadpour., Community aware random walk for network embedding, Knowledge-Based Systems 148, 2018, p 47-54

https://www.sciencedirect.com/science/article/pii/S0950705118300911

[43] Sun, Yizhou, et al., Pathsim: Meta path-based top-k similarity search in heterogeneous information networks, Proceedings of the VLDB Endowment 4.11, 2011, p 992-1003

https://dl.acm.org/doi/pdf/10.14778/3402707.3402736

[44] Zhang, Chuxu, et al., Heterogeneous graph neural network, Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019

https://dl.acm.org/doi/pdf/10.1145/3292500.3330961

[45] Scarselli, Franco, et al., The graph neural network model, IEEE transactions on neural networks 20.1, 2008, p 61-80

https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4700287

[46] Kipf, Thomas N., and Max Welling., Semi-supervised classification with graph convolutional networks, arXiv preprint arXiv:1609.02907, 2016

https://arxiv.org/pdf/1609.02907.pdf

[47] Newman, Mark EJ., Clustering and preferential attachment in growing networks, Physical review E 64.2, 2001, 025102

https://arxiv.org/pdf/cond-mat/0104209.pdf

[48] De Sá, Hially Rodrigues, and Ricardo BC Prudêncio., Supervised link prediction in weighted networks, The 2011 international joint conference on neural networks. IEEE, 2011

https://cin.ufpe.br/~rbcp/papers/ijcnn-2011.pdf

[49] Jaccard, Paul., Distribution comparée de la flore alpine dans quelques régions des Alpes occidentales et orientale, Bulletin de la Murithienne 31, 1902, p 81-92

https://doc.rero.ch/record/23401/files/BCV_N_112_031_1902_081.pdf

[50] Wagner, Caroline S., and Loet Leydesdorff., Mapping the network of global science: comparing international co-authorships from 1990 to 2000, International journal of Technology and Globalisation 1.2, 2005, p 185-208

https://www.leydesdorff.net/sciencenets/mapping.pdf

[51] Ravasz, Erzsébet, et al., Hierarchical organization of modularity in metabolic networks, science 297.5586, 2002, p 1551-1555

https://arxiv.org/ftp/cond-mat/papers/0209/0209244.pdf

[52] Lü, Linyuan, and Tao Zhou, Link prediction in complex networks: A survey, Physica A: statistical mechanics and its applications 390.6, 2011, p 1150-1170

https://www.sciencedirect.com/science/article/pii/S037843711000991X

[53] Leicht, Elizabeth A., Petter Holme, and Mark EJ Newman, Vertex similarity in networks, Physical Review E 73.2, 2006, 026120

https://arxiv.org/pdf/physics/0510143.pdf

[54] Sun, Qingshuang, et al, An improved link prediction algorithm based on degrees and similarities of nodes, 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS). IEEE, 2017

https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7959962

[55] Adamic, Lada A., and Eytan Adar, Friends and neighbors on the web, Social networks 25.3 (2003): 211-230.

https://www.sciencedirect.com/science/article/pii/S0378873303000091

[56] Zhou, Tao, Linyuan Lü, and Yi-Cheng Zhang, Predicting missing links via local information, The European Physical Journal B 71.4, 2009, p 623-630

https://link.springer.com/content/pdf/10.1140/epjb/e2009-00335-8.pdf

[57] Barabási, Albert-László, and Réka Albert, Emergence of scaling in random networks, science 286.5439, 1999, p 509-512

https://arxiv.org/pdf/cond-mat/9910332.pdf

[58] Pearson, Karl, Notes on regression and inheritance in the case of two parents, Proceedings of the Royal Society of London. 58, 1895, p 240–242

https://books.google.sk/books?id=60aL0zlT-90C&pg=PA240

[59] Spearman, C.,"General Intelligence," Objectively Determined and Measured, The American Journal of Psychology. 15 (2), 1904, p 201–292

https://www.jstor.org/stable/1412107

[60] Kendall, Maurice G., A new measure of rank correlation, Biometrika 30.1/2, 1938, p 81-93

https://www.jstor.org/stable/pdf/2332226.pdf

[61] KIM ESBENSEN, PAUL GELADI, Principal Component Analysis, Chemometrics and Intelligent Laboratory Systems, 2, 1987, p 37-52

http://pzs.dstu.dp.ua/DataMining/pca/bibl/Principal%20components%20analysis.pdf

[62] Mardani, Abbas, et al., Multiple criteria decision-making techniques and their applications–a review of the literature from 2000 to 2014, Economic research-Ekonomska istraživanja 28.1, 2015, p 516-571

https://www.tandfonline.com/doi/pdf/10.1080/1331677X.2015.1075139

[63] Hwang, Ching-Lai, Young-Jou Lai, and Ting-Yun Liu, A new approach for multiple objective decision making, Computers & operations research 20.8, 1993, p 889-899

https://www.sciencedirect.com/science/article/abs/pii/030505489390109V

[64] Girvan, Michelle, and Mark EJ Newman, Community structure in social and biological networks, Proceedings of the national academy of sciences 99.12, 2002, p 7821-7826

https://www.pnas.org/content/pnas/99/12/7821.full.pdf

[65] Pons, Pascal, and Matthieu Latapy, Computing communities in large networks using random walks, International symposium on computer and information sciences. Springer, Berlin, Heidelberg, 2005

https://www-complexnetworks.lip6.fr/~latapy/Publis/communities.pdf

[66] Ward Jr, Joe H., Hierarchical grouping to optimize an objective function, Journal of the American statistical association 58.301, 1963, p 236-244

https://www.jstor.org/stable/pdf/2282967.pdf

[67] Vinh, Nguyen Xuan, Julien Epps, and James Bailey, Information theoretic measures for clusterings comparison: is a correction for chance necessary?, Proceedings of the 26th annual international conference on machine learning, 2009

https://dl.acm.org/doi/pdf/10.1145/1553374.1553511