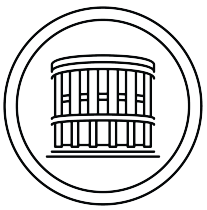
UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



VPLYV DYNAMICKÉHO REFORMÁTOVANIA SIETE SOCIÁLNYCH KONTAKTOV PRI ŠÍRENÍ INFEKČNÝCH CHORÔB

DIPLOMOVÁ PRÁCA

2024 Bc. Peter ONDRUŠEK

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

VPLYV DYNAMICKÉHO REFORMÁTOVANIA SIETE SOCIÁLNYCH KONTAKTOV PRI ŠÍRENÍ INFEKČNÝCH CHORÔB

DIPLOMOVÁ PRÁCA

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

Bratislava 2024 **Bc. Peter Ondrušek**



67423296

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Peter Ondrušek
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: Vplyv dynamického reformátovania siete sociálnych kontaktov pri šírení infekčných chorôb
Influence of dynamical reformatting of social contact networks on infection disease spreading

Anotácia: Mnohé mechanizmy a vlastnosti prebiehajúcej pandémie dobre popisujú matematické modely šírenia infekcie na náhodných sieťach s mocninovo rozdeleným stupňom vrcholov. V týchto modeloch možno simulovať šírenie infekcie aj vakcináciu jednotlivcov, ktoré vedú k zmenám siete sociálnych kontaktov. Avšak sieť sociálnych kontaktov a intenzita kontaktov sa mení počas pandémie dynamicky aj sama na základe zmeny správania jednotlivcov. Cieľom diplomovej práce je študovať priebeh simulovanej epidémie v dynamických reformátujúcich grafoch, špeciálne dynamiku analógu epidemiologického parametra reprodukčného čísla a otázku dosiahnutia kolektívnej imunity pri infekcii. Výsledky práce môžu mať priame implikácie na vysvetlenie pozorovanej dynamiky pandémie SARS-CoV-2.

Literatúra: Mark Newman - Networks (2016)

Vedúci: doc. Mgr. Richard Kollár, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: doc. Mgr. Igor Melicherčík, PhD.

Dátum zadania: 05.01.2023

Dátum schválenia: 11.01.2023 prof. RNDr. Daniel Ševčovič, DrSc.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie Touto cestou by som sa chcel poďakovať svojmu vedúcemu diplomovej práce doc. Mgr. Richardovi Kollárovi, PhD. za pomoc, ochotu, odborné rady, motiváciu a konzultácie počas celej doby písania práce. Tiež ďakujem svojej rodine a priateľom za ich podporu a trpezlivosť.

Abstrakt

ONDRUŠEK, Peter: Vplyv dynamického reformátovania siete sociálnych kontaktov pri šírení infekčných chorôb [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: doc. Mgr. Richard Kollár, PhD., Bratislava, 2024, 56 s.

Predmetom skúmania našej diplomovej práce je študovať priebeh simulovanej epidémie v dynamickom reformátujúcom grafe. V úvodnej kapitole je predstavený jeden zo základných matematických modelov využívaných pri štúdiách šírenia infekčných chorôb – SIR model. Zavedený je pojem reprodukčného čísla. Následne sú uvedené základné vlastnosti sietí a objasnený pojem sociálna sieť. Tretia kapitola je venovaná základom teórie grafov. Vysvetlené sú základné vlastnosti matice susedností a predstavené rôzne typy grafov. Tiež je predstavený konfiguračný model, ako jeden zo základných algoritmov používaných pri generovaní náhodných grafov. Štvrtá kapitola sa venuje nami vytvorenému modelu, ktorý pozostáva z dvoch častí. Prvou je vytvorenie grafu na základe určitých charakteristík a druhou časťou je samotná simulácia epidémie na vytvorenom grafe. Na záver sú prezentované príklady realizácií simulácií a analýza získaných výsledkov.

Kľúčové slová: Reprodukčné číslo, Graf, Sieť, SIR model, Konfiguračný model.

Abstract

ONDRUŠEK, Peter: Influence of dynamical reformatting of social contact networks on infection disease spreading [Bachelor Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: doc. Mgr. Richard Kollár, PhD., Bratislava, 2022, 56p.

The subject of our thesis is to study the course of a simulated epidemic on a dynamic reformatting graph. In the introductory chapter, one of the basic mathematical models used in studies of the spread of infectious diseases - the SIR model - is presented. The concept of reproduction number is introduced. Subsequently, the basic characteristics of networks and the term social network are clarified. The third chapter is dedicated to the basics of graph theory. The basic properties of the adjacency matrix are explained and different types of graphs are presented. The configuration model is also presented as one of the basic algorithms used in the generation of random graphs. The fourth chapter deals with the model we created, which consists of two parts. The first is the creation of a graph based on certain characteristics, and the second part is the actual simulation of the epidemic on the created graph. At the end, examples of simulations and analysis of the obtained results are presented.

Keywords: Reproduction number, Graph, Network, SIR model, Configuration model.

Obsah

Úvod **8**

1 SIR model **9**

2 Úvod do teórie sietí **11**

2.1 Sociálne siete 13

3 Teória grafov **14**

3.1 Konfiguračný model 16

4 Model simulácie epidémie **18**

4.1 Vytvorenie siete 18

4.2 Simulácia epidémie 27

5 Vyhodnocovanie modelu **33**

5.1 Realizácie pozorovaní 34

5.2 Analýza grafov 37

5.2.1 Vzťah kriviek zmien počtu nakazených 37

5.2.2 Vzťah zmeny počtu nakazených a maximálnej vlastnej hodnoty . . . 45

5.3 Tri fázy priebehu epidémie 52

Záver **54**

Zoznam použitej literatúry **56**

Úvod

V súčasnej dobe globalizácie, kde sa informácie šíria rýchlejšie než kedykoľvek predtým, má dynamika sociálnych sietí zásadný vplyv na šírenie infekčných chorôb. Táto diplomová práca sa zameriava na preskúmanie vplyvu dynamického reformátovania sietí sociálnych kontaktov na šírenie infekčných chorôb. Sociálne siete poskytujú rámec pre vzájomné interakcie jednotlivcov, a preto je štúdium ich dynamiky a štruktúry kľúčové pre porozumenie a riadenie epidemických javov.

V prvej kapitole si v krátkosti predstavíme jeden zo základných matematických modelov používaných pri simuláciách a štúdiách šírenia infekčných chorôb. Popíšeme si matematickú reprezentáciu tohto modelu a zavedieme s tým spojený pojem reprodukčného čísla.

Druhá kapitola tejto práce predstavuje základný teoretický rámec sietí, kde sú vysvetľované kľúčové pojmy. Uvedieme základné delenie sietí a príklady k nim. Základnou metrikou v teórii sietí je stupeň vrchola, ktorý nám hovorí o prepojenosti s ostatnými vrcholmi.

Nasledujúca kapitola sa hlbšie venuje teórii grafov a predstavuje ich základné tipy. Približujeme matematickú reprezentáciu grafov, ktorá bude kľúčová pre našu diplomovú prácu. V závere tejto kapitoly uvedieme jeden zo základných algoritmov pre generovanie grafov, konfiguračný model, ktorý poskytuje dôležitý nástroj pre modelovanie štruktúry sociálnych sietí.

Vo štvrtej kapitole predstavíme nami vytvorený model simulácie epidémie vyvinutý v prostredí R [6]. Jeho prvou časťou je vytvorenie siete ľudských kontaktov, s dôrazom na čo najvernejšie priblíženie realite. Pri jej tvorbe budeme využívať práve konfiguračný model a ďalšie procesy vysvetľované v tretej kapitole. V druhej časti si dôkladne popíšeme princípy simulácie priebehu epidémie na nami vytvorenej sieti.

V poslednej kapitole sa pozrieme na konkrétne realizácie našich pozorovaní pre rôzne nastavenia parametrov modelu. Budeme sa zaoberať dôkladnou analýzou získaných výsledkov. Na základe nich odvodíme užitočné informácie o priebehu epidémie v dynamických reformátujúcich grafoch.

1 SIR model

V úvodnej kapitole si v krátkosti predstavíme jeden zo základných epidemiologických modelov, pričom čerpať budeme najmä z [4] a [1]. Uvedieme iba základné charakteristiky, keďže skúmanie modelu nie je cieľom našej diplomovej práce.

SIR model je matematický model, ktorý sa používa na simuláciu a štúdium šírenia infekčných chorôb, ako sú napríklad chrípka, tuberkulóza alebo aj nedávne ochorenia ako COVID-19. Tento model rozdeľuje populáciu do troch základných skupín:

- **S (Susceptible):** vnímaví jedinci, ktorí sú zdraví a daným ochorením sa môžu nakaziť,
- **I (Infected):** infekční jedinci, ktorí sú ochorením nakazení a ochorenie môžu ďalej šíriť kontaktom s vnímavým jedincom,
- **R (Recovered):** vyliečení jedinci, ktorí chorobu už prekonali alebo sú imúnni voči nej.

Pri ochoreniach často dochádza aj k úmrtiam. Títo zomrelí jedinci sa zaraďujú do tretej skupiny, keďže v tomto modeli zohrávajú rovnakú úlohu. Základný SIR model je reprezentovaný tromi diferenciálnymi rovnicami:

$$\begin{aligned} \dot{S} &= -\beta SI \\ \dot{I} &= \beta SI - \gamma I \\ \dot{R} &= \gamma I, \end{aligned} \tag{1}$$

kde S, I a R sú normované počty jedincov danej skupiny, t. j. spĺňajú rovnicu $S+I+R = 1$. Hodnota β vyjadruje pravdepodobnosť prenosu infekcie od infikovaného jedinca na vnímavého. Parameter γ vyjadruje dĺžku infekčnosti jedinca, t. j. aký dlhý čas môže infekčný jedinec nakaziť vnímavého jedinca. Tento parameter môžeme vyjadriť ako $\gamma = \frac{1}{T}$, kde T je priemerný čas, za ktorý infekčný jedinec infikuje vnímavého jedinca. V našom modeli budeme mať nastavené $T = 1$ (deň), t. j. aj parameter $\gamma = 1$. To znamená, že infekčný jedinec môže nakaziť vnímavých jedincov zo svojich kontaktov iba počas jedného kroku (dňa) priebehu epidémie.

Druhú rovnicu modelu (1) vieme upraviť nasledovne:

$$\dot{I} = \beta SI - \gamma I = (\beta S - \gamma)\, I = \gamma I \left(\frac{\beta S}{\gamma} - 1 \right),$$

kde výraz $\frac{\beta S}{\gamma} = R(t)$ vyjadruje efektívne reprodukčné číslo. Teda po diskretizácii tejto rovnice získame:

$$\begin{aligned}
\dot{I} &= \gamma\,(R(t) - 1)\, I \\
\frac{I_{n+1} - I_n}{\Delta t} &\approx \gamma\,(R(t) - 1)\, I_n \\
I_{n+1} - I_n &\approx \gamma \Delta t\,(R(t) - 1)\, I_n \\
I_{n+1} - I_n &\approx R(t) I_n - I_n \\
I_{n+1} &\approx R(t) I_n \\
\frac{I_{n+1}}{I_n} &\approx R(t).
\end{aligned}$$

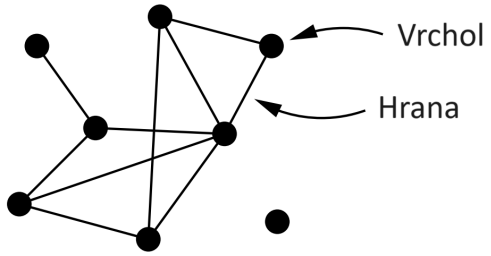
Keďže v našom modeli budeme uvažovať parameter $\gamma = 1$ a taktiež dĺžku kroku $\Delta t = 1$, vyššie znázorneným postupom vidíme, že efektívne reprodukčné číslo vieme aproximatívne vyjadriť ako zmenu počtu nakazených v čase.

Typicky sa identifikuje reprodukčné číslo v SIR modeli ako priemerný počet infikovaných jedincov od jedného infekčného jedinca, avšak práve cieľom tejto práce je lepšie objasniť, aký praktický význam v pozorovaných dátach ma reprodukčné číslo v modeloch, ktoré vernejšie modelujú šírenie infekcie v populácii. Je to základný ukazovateľ potenciálu šírenia infekčného ochorenia a hovorí o celkovom potenciáli samotnej epidémie. Rozlišujeme základné reprodukčné číslo, čo je iba teoretická hodnota a je to časovo nemenná vlastnosť infekcie. Naopak, efektívne reprodukčné číslo nám hovorí, ako rýchlo sa mení počet infikovaných v čase, v závislosti od aktuálnej situácie.

2 Úvod do teórie sietí

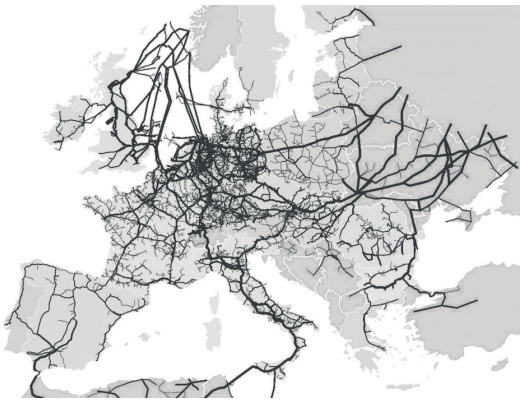
Teória v tejto kapitole je spracovaná podľa [4] a [5].

Jednoducho povedané, sieť je súbor bodov pospájaných čiarami do párov. V teórii sietí tieto body označujeme ako uzly alebo vrcholy, a čiary ako hrany (obr. 1).

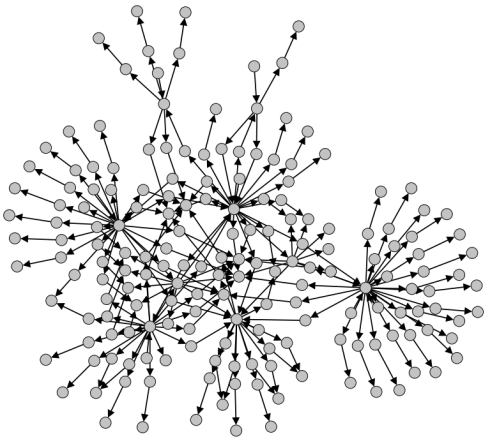


Obr. 1: Základné prvky siete [4, str. 1].

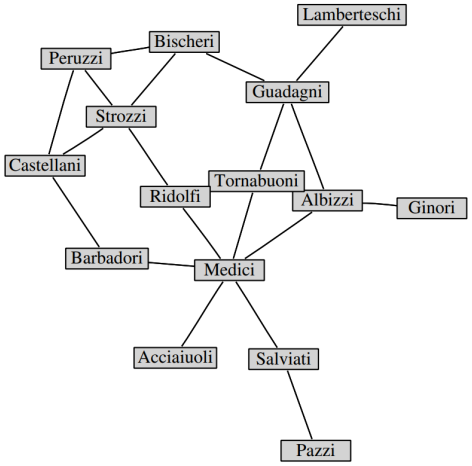
Siete rôznych typov vznikajú takmer vo všetkých odvetviach vedy a techniky. Môžeme ich rozdeliť do 4 základných kategórií, a to technologické siete (počítačová dátová sieť, sieť ciest, či železničných tratí, obr. 2), informačné siete (sieť webových stránok, obr. 3), sociálne siete (vrcholy sú ľudia a hrany spojenia medzi nimi, obr. 4) a biologické siete (neurónové siete, potravinové reťazce, obr. 5).



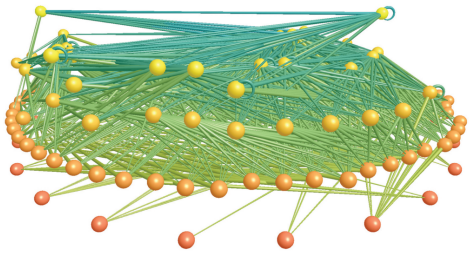
Obr. 2: Technologická sieť - sieť plynovodov v Európe [4, str. 30].



Obr. 3: Informačná sieť - sieť webových stránok [4, str. 33].



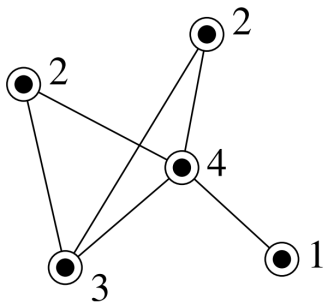
Obr. 4: Sociálna sieť - sieť zmiešaných manželstiev rodov Florencie v 15. storočí [4, str. 59].



Obr. 5: Biologická sieť - potravinová sieť jazera Little Rock Lake vo Wisconsine [4, str. 6].

Siete delíme ďalej na orientované (napr. webové stránky, stránka A môže odkazovať na stránku B, ale stránka B nemusí odkazovať na stránku A) a neorientované (sieť sociálnych kontaktov, ak je osoba A v kontakte s osobou B, tak je aj osoba B v kontakte s osobou A).

Pre pochopenie významu sietí používame rôzne metriky. Najjednoduchšou je stupeň vrchola, ktorý je určený počtom hrán, ktoré sú k nemu pripojené (obr. 6). V neorientovaných sieťach je to jedno číslo, v našom prípade počet ľudí, s ktorými je daný človek (vrchol) v kontakte. V orientovaných sieťach rozlišujeme *in-degree* a *out-degree*, počet hrán, ktoré vstupujú do daného vrchola, resp. počet hrán, ktoré z neho vychádzajú.



Obr. 6: Sieť so znázornenými stupňami jednotlivých vrcholov [4, str. 9].

2.1 Sociálne siete

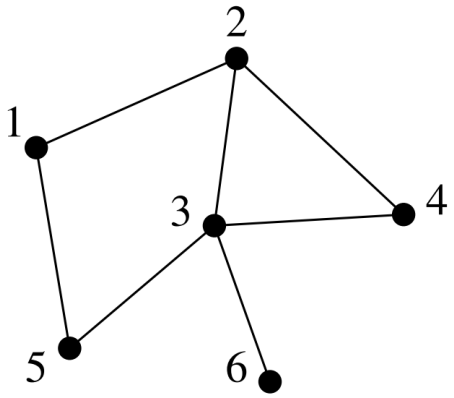
Pod pojmom sociálna sieť si každý z nás ako prvé predstaví sociálne médiá ako Facebook, Instagram, či Twitter. V skutočnosti má pojem sociálna sieť oveľa širší význam a predstavuje ľubovoľnú sieť, v ktorej vrcholmi sú ľudia a hrany predstavujú určitú formu prepojenia medzi nimi. Od formy tohto prepojenia závisí, aké otázky nás budú zaujímať. Hrany môžu predstavovať priateľstvo medzi ľuďmi, ich profesionálne vzťahy, výmenu tovaru či peňazí medzi nimi, romantické či sexuálne vzťahy a mnohé ďalšie.

Napríklad sieť priateľstiev na Facebooku môžeme považovať za neorientovanú sieť (ak je používateľ A priateľom používateľa B, tak je aj používateľ B priateľom používateľa A). Na Instagrame ide naopak o orientovanú sieť (používateľ A sleduje používateľa B, ale B nemusí sledovať používateľa A).

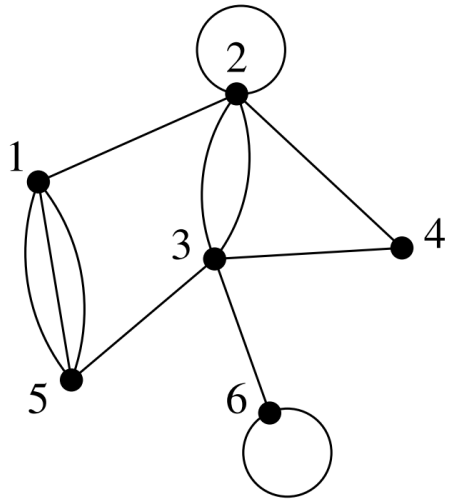
3 Teória grafov

Teória v tejto kapitole je spracovaná podľa [4].

Chceli by sme siete matematicky analyzovať, na čo používame základy teórie grafov, keďže každú sieť môžeme tiež definovať ako graf. Každý graf je definovaný počtom vrcholov, ktorý budeme značiť N. Počet hrán definujeme ako m. Väčšina grafov má jednoduché hrany medzi dvojicou vrcholov (obr. 7). Môže však nastať situácia, kedy medzi dvoma vrcholmi je multihrana, resp. viacnásobná hrana (medzi vrcholmi 1 a 5 aj vrcholmi 2 a 3, obr. 8) alebo vlastná hrana (vrchol 2 aj 6, obr. 8), ktorá spája bod sám so sebou. My budeme pracovať s jednoduchými grafmi, teda grafmi bez multihrán a vlastných hrán.



Obr. 7: Jednoduchý graf - iba jednoduché hrany [4, str. 107].



Obr. 8: Multigraf - multihrany aj vlastné hrany [4, str. 107].

Základnou matematickou reprezentáciou grafu je matica susedností A. Hovorí nám, ktoré vrcholy sú navzájom prepojené, teda:

$$A_{ij} = \begin{cases} 1, & \text{ak je hrana medzi bodmi } \quad i \quad \text{a} \quad j, \\ 0, & \text{inak.} \end{cases}$$

Matica susedností pre neorientované grafy je vždy symetrická, pre jednoduché grafy má nulovú diagonálu a mimo diagonály iba 0 alebo 1.

Matica susedností grafu na obr. 7 by bola:

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

Pre multigraf nemusí mať matica nulovú diagonálu (vlastné hrany) a mimo diagonály aj celé čísla väčšie ako 1 (multihrany). Matica susedností grafu na obr. 8 by bola:

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 3 & 0 \\ 1 & 2 & 2 & 1 & 0 & 0 \\ 0 & 2 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 3 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 2 \end{pmatrix}.$$

V niektorých prípadoch chceme mať vážené hrany, ktoré nám hovoria o intenzite spojenia medzi dvomi vrcholmi. Takýto graf nazývame vážený graf. V matici susedností takéhoto grafu sa mimo diagonály nachádzajú hodnoty z intervalu $(0, 1)$.

Z matice susedností vieme určiť stupeň vrchola k_i ako súčet:

$$k_i = \sum_{j=1}^{n} A_{ij}.$$

Potom vieme určiť aj jednu zo základných metrík pre analýzu grafov, a to priemerný stupeň vrcholov ako:

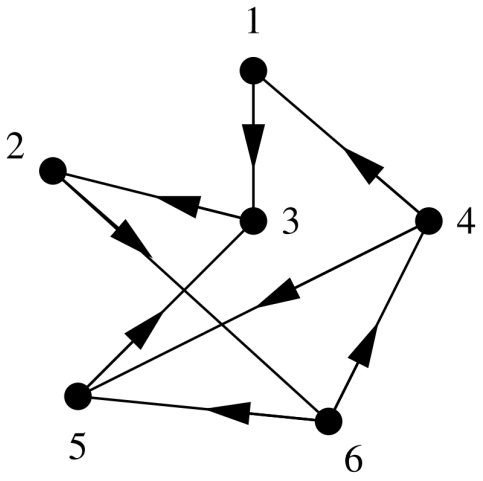
$$c = \frac{1}{n} \sum_{i=1}^{n} k_i.$$

Pre orientované grafy matica nespĺňa vlastnosť symetrie a je definovaná ako:

$$A_{ij} = \begin{cases} 1, & \text{ak existuje hrana z vrcholu } \quad j \quad \text{do vrcholu } \quad i, \\ 0, & \text{inak.} \end{cases}$$

Matica susedností pre orientovaný graf na obr. 9 by bola nasledovná:

$$A = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}.$$



Obr. 9: Orientovaný graf.

Ďalšími metrikami sú tzv. centrality, ktoré nám hovoria o dôležitosti daného vrcholu v grafe. Poznáme napríklad stupňovú centralitu, centralitu vlastného vektora, Katzovu centralitu, centralitu blízkosti či prepojenosti a iné.

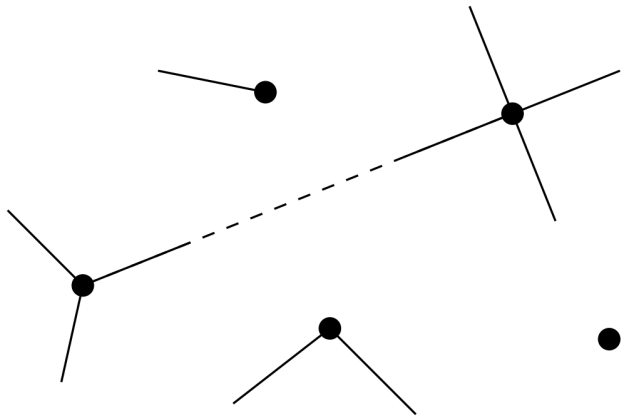
3.1 Konfiguračný model

Existuje množstvo rôznych algoritmov na generovanie náhodných grafov, spomedzi ktorých je konfiguračný model ideálnym prepojením jednoduchosti a obrazom reálneho sveta. Je jedným z najdôležitejších teoretických modelov používaných pri štúdiu sietí.

Konfiguračný model je modelom náhodného grafu s danou postupnosťou stupňov vrcholov. Poznáme teda presný stupeň každého jedného vrcholu v sieti. Aby bolo možné graf vygenerovať, ako už vieme z teórie grafov, dôležité je, aby bol súčet stupňov v danej postupnosti párny.

Uvažujme teda, že poznáme stupeň k_i každého vrcholu $i = 1, \ldots, N$ a že súčet týchto stupňov $\sum_{i=1}^{N} k_i$ je párny. Každému vrcholu i pridelíme k_i „výbežkov“, tiež nazývaných polhrany. Dostávame celkovo $\sum_{i=1}^{N} k_i = 2m$ výbežkov, kde m bude počet hrán vo vygenerovanom grafe. Proces generovania náhodného grafu spočíva v tom, že náhodne vyberieme dva výbežky, ktorých spojením vytvoríme hranu (ako je znázornené na obr. 10). Následne spojíme ďalšiu dvojicu zo zvyšných $2m - 2$ výbežkov a pokračujeme, pokiaľ nebudú navzájom poprepájané všetky výbežky. Tento proces vytvorí konkrétne párovanie výbežkov,

pričom generuje každé možné párovanie s rovnakou pravdepodobnosťou.



Obr. 10: Proces spájania dvoch polhrán k vytvoreniu hrany [4, str. 370].

Nevýhodou pri takto generovaných grafoch je vznik vlastných hrán a multihrán. Keďže dvojicu výbežkov vyberáme náhodne, nevieme zabrániť tomu, aby sa spojili dva výbežky vychádzajúce z rovnakého vrcholu alebo výbežky dvoch vrcholov, medzi ktorými už hrana existuje. Dalo by sa tomu vyhnúť počas procesu generovania zamietnutím vytvorenia takejto hrany, avšak takto vygenerované siete sa nevytvárajú rovnomerne z množiny všetkých možných párovaní. To by pokazilo určité vlastnosti siete, ktorú chceme neskôr analyzovať. Tiež by mohla nastať situácia, kedy by sa proces zasekol. Predstavme si situáciu, kde nám zostanú posledné dva výbežky, ktoré vychádzajú z toho istého vrcholu alebo patria vrcholom, medzi ktorými už hrana existuje. V takomto prípade by proces zlyhal.

Niektoré siete v reálnom svete môžu obsahovať multihrany alebo vlastné hrany , no väčšina ich nemá. Avšak keď pracujeme s veľkými sieťami (veľký počet vrcholov), ukazuje sa, že hustota multihrán a vlastných hrán v konfiguračnom modeli sa blíži k nule, čiže ich vplyv pri našej analýze bude zanedbateľný.

Pri inicializácii procesu sme uvažovali známu postupnosť stupňov vrcholov. Často však chceme pracovať so štatistickou distribúciou stupňov. To vieme jednoducho vyriešiť tým, že zo zadanej distribúcie vykreslíme postupnosť stupňov a ďalej pokračujeme v procese spomínanom vyššie. Daná štatistická distribúcia pridelí každému možnému stupňu k pravdepodobnosť jeho výskytu p_k. Pri štúdii sociálnych sietí sa používa mocninové rozdelenie stupňov.

4 Model simulácie epidémie

V tejto kapitole predstavíme nami vytvorený model simulácie epidémie (v softvéri R [6]), ktorý pozostáva z dvoch hlavných častí. Prvou časťou je vytvorenie siete ľudských kontaktov na základe určitých parametrov a charakteristík pomocou teórie vysvetľovanej v predchádzajúcej kapitole. Druhou časťou bude samotná simulácia epidémie na danej vytvorenej sieti.

4.1 Vytvorenie siete

V tejto časti budeme generovať sieť ľudských kontaktov pomocou niekoľkých parametrov a charakteristík tak, aby sa čo najviac približovala realite. Základným parametrom pri generovaní bude počet vrcholov N, t. j. počet ľudí v našej sieti. Chceli by sme N čo najväčšie, no zároveň výpočtovo zvládnuteľné. Pri analýze budeme najčastejšie používať $N = 500$.

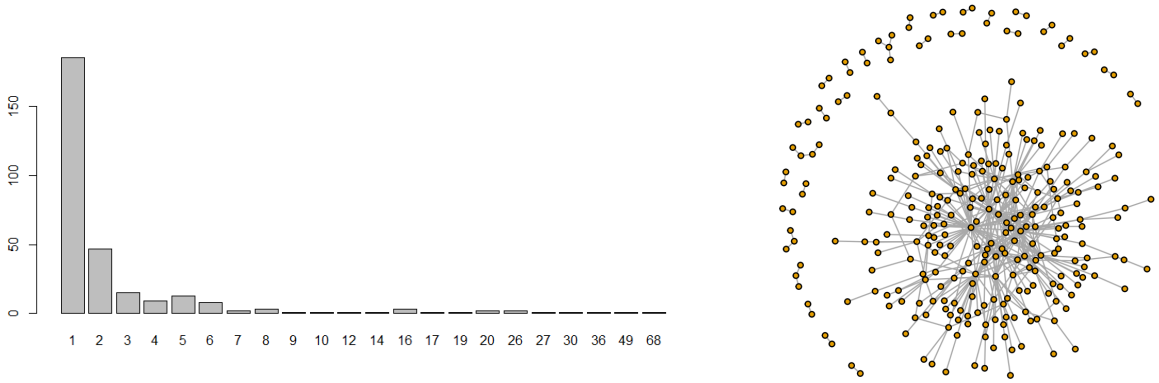
Mocninové rozdelenie

Ďalší je parameter mocninového rozdelenia stupňov vrcholov α. Pri vytváraní postupnosti stupňov vrcholov z tejto distribúcie si vytvoríme pomyselnú mriežku, ktorej dieliky pre stupne $k = 1, 2, \ldots, N$ budú o veľkosti $\frac{1}{k^\alpha}$, teda $1, \frac{1}{2^\alpha}, \ldots, \frac{1}{N^\alpha}$. Ak si označíme sumu veľkostí týchto dielikov ako $s_N = \sum_{k=1}^{N} \frac{1}{k^\alpha}$, tak potom pravdepodobnosť výskytu vrcholu so stupňom k bude:

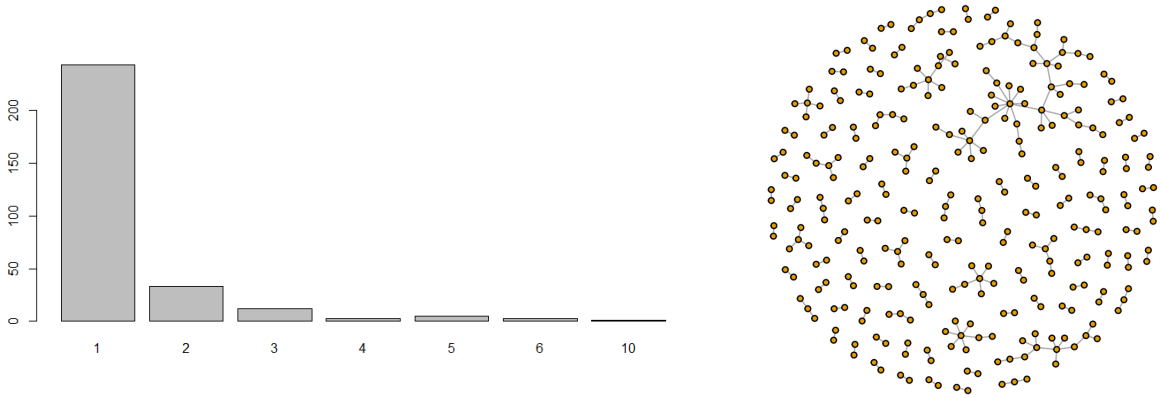
$$p_k = \frac{\frac{1}{k^\alpha}}{s_N}.$$

Hodnoty α sa typicky pohybujú v intervale $2 < \alpha < 3$. Čím viac sa hodnota tohto parametru blíži k číslu 2, tým je postupnosť stupňov rozsiahlejšia, t. j. obsahuje aj vrcholy s vyššími stupňami, a preto je aj príslušná sieť hustejšia (viď obr. 11 pre $\alpha = 2,1$). Na obrázku vidíme, že sieť sa skladá z jedného veľkého komponentu súvislostí a niekoľkých malých komponentov. Naopak, ak sa hodnota parametru α blíži k číslu 3, tým sa dieliky pre zvyšujúce sa k zmenšujú v pomere k dielikom s nižším stupňom, t. j. pravdepodobnosť p_k výskytu stupňa vrchola k sa pre zvyšujúce k znižuje. Dostávame postupnosť s nízkymi stupňami vrcholov, a tým pádom aj oveľa redšiu príslušnú sieť (viď obr. 12 pre $\alpha = 2,9$).

Vidíme, že stupeň 1 tvorí až okolo 80% danej postupnosti a najvyšší stupeň je 10, oproti prvému príkladu, kedy sme dostali viacero vrcholov s vyšším stupňom ako 10. V danej sieti sa nenachádza žiadny veľký komponent a skladá sa z veľkého počtu malých komponentov. Ukazuje sa, že pri analýze sociálnych sietí je ideálne nastavenie parametra na hodnotu približne $\alpha = 2.4$ až $\alpha = 2.5$, kedy sa sieť najlepšie približuje realite.



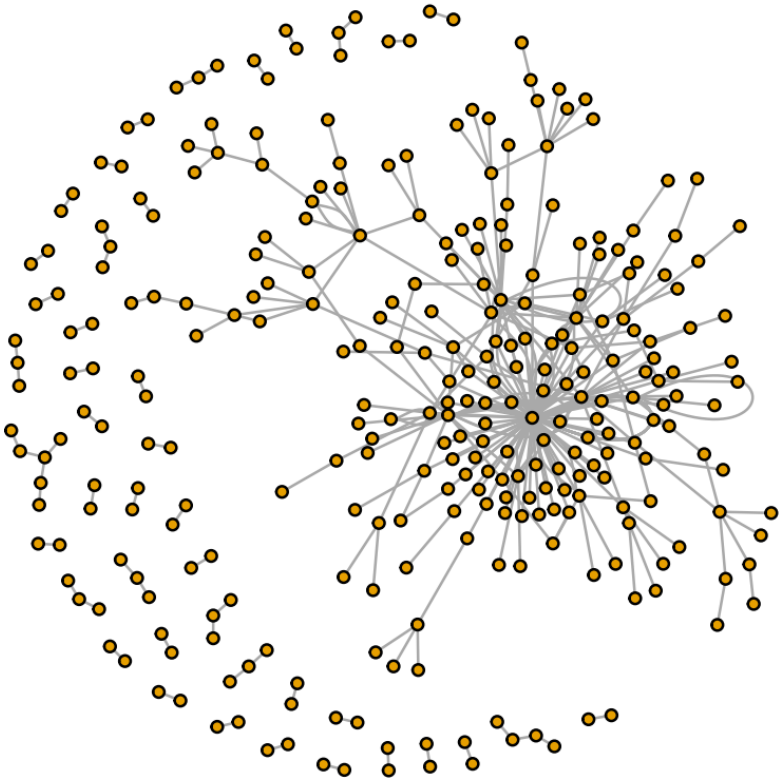
Obr. 11: Histogram stupňov vrcholov (vľavo) a príslušná sieť (vpravo) pre $N = 300$ a $\alpha = 2,1$ (vlastné spracovanie v [6]).



Obr. 12: Histogram stupňov vrcholov (vľavo) a príslušná sieť (vpravo) pre $N = 300$ a $\alpha = 2,9$ (vlastné spracovanie v [6]).

Implementácia mocninového rozdelenia

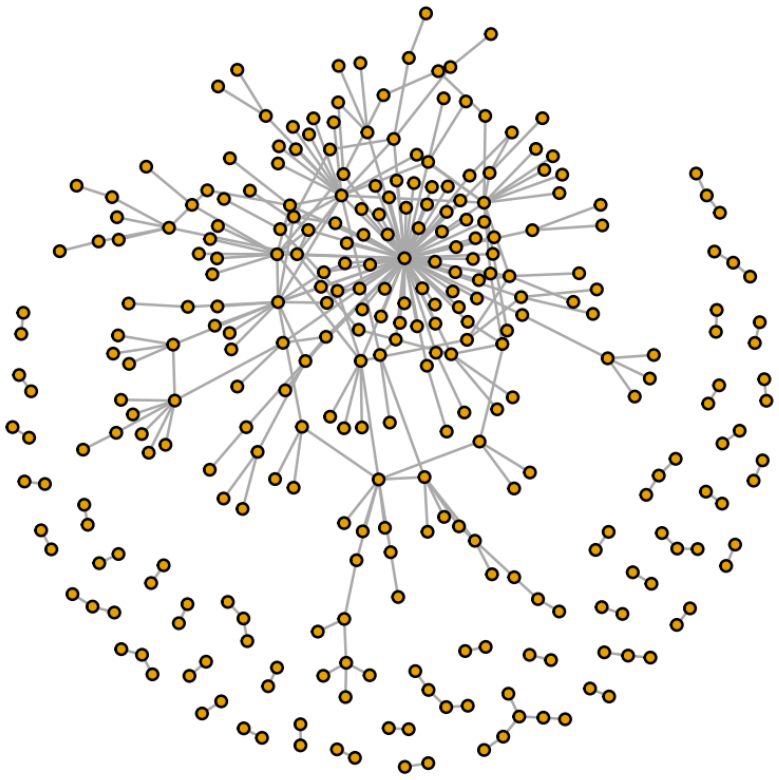
V softvéri R (ďalej R-ko) si vygenerujeme postupnosť stupňov vrcholov prislúchajúcu mocninovému rozdeleniu s príslušnými parametrami N a α pomocou funkcie *rpldis*. Na vytvorenie grafu z danej postupnosti stupňov využijeme konfiguračný model, ktorého princíp fungovania sme opísali v kapitole 3.1. Počiatočnou podmienkou k vytvoreniu grafu je párny súčet všetkých stupňov vrcholov. Túto podmienku overíme a ak je súčet nepárny, náhodne vyberieme jeden bod z postupnosti a zvýšime mu hodnotu o 1. Všetko potrebné pre použitie konfiguračného modelu máme pripravené. V R-ku vygenerujeme takýto graf pomocou funkcie *sample_ degseq*. Ako však vieme, graf vygenerovaný pomocou konfiguračného modelu môže obsahovať aj vlastné hrany a multihrany (viď obr. 13).



Obr. 13: Vygenerovaný graf pomocou konfiguračného modelu obsahujúci vlastné hrany aj multihrany (vlastné spracovanie v [6]).

Vlastné hrany a multihrany

V reálnej sieti ľudských kontaktov nemá zmysel uvažovať, že je niekto v kontakte sám so sebou (vlastná hrana) alebo majú dvaja ľudia viac kontaktov (multihrana). Tento nedostatok vieme jednoducho vyriešiť použitím matice susedností danej siete. Hodnoty na diagonále prepíšeme na nulu, čím odstránime vlastné hrany. V druhom kroku sa pozrieme na pozície mimo diagonály a všetky hodnoty väčšie ako jedna, ktoré nám predstavujú multihrany, zmeníme na jednotky, čím z multihrán dostávame jednoduché hrany. Po týchto úpravách vygenerujeme podľa upravenej matice susedností príslušný graf pomocou funkcie *graph.adjacency*. Dostávame teda jednoduchý graf (viď obr. 14).



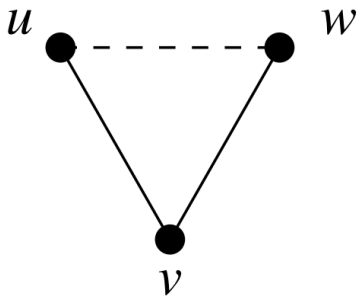
Obr. 14: Upravený graf po odstánení vlastných hrán a multihrán (vlastné spracovanie v [6]).

Odstránením vlastných hrán a multihrán neporušíme základné vlastnosti siete, kedže ich hustota sa v konfiguračnom modeli pre dostatočne veľké N blíži k nule. Mocninové rozdelenie stupňov vrcholov sa taktiež zachováva po týchto úpravách.

Dopĺňanie hrán do trojuholníka

Dostali sme sieť, ktorá sa už približuje realite, no stále je pomerne riedka a roztrúsená. Chceli by sme vytvoriť akési skupinky ľudí, kotré by predstavovali rodiny žijúce v jednej domácnosti, žiakov v škole, športovcov v klube a pod. Budeme predpokladať, že ak určitý človek (vrchol v grafe) je v kontakte (prepojený hranami) s dvomi ďalšími ľuďmi (vrcholmi v grafe), tak je veľká pravdepodobnosť, že aj tieto dve osoby sú navzájom v kontakte. Ak medzi nimi hrana nie je, cheli by sme ju s určitou pravdepodobnosťou doplniť a vytvoriť tak trojuholník (viď obr. 15).

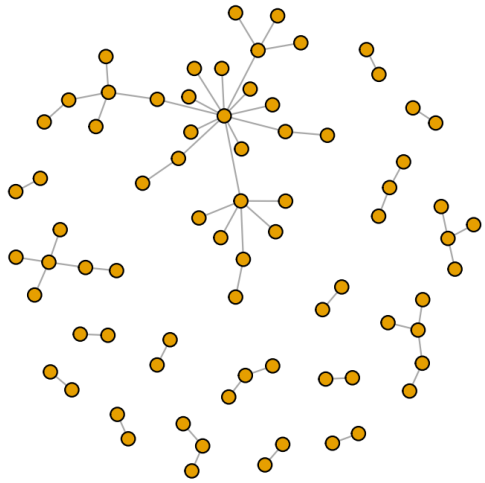
V našom programe najprv náhodne vyberieme jednu hranu zo zoznamu všetkých hrán grafu (na ilustračnom príklade na obr. 15 napr. hrana spájajúca body u a v). Následne si náhodne zvolíme jeden z bodov (v príklade bod v), ktoré daná hrana spája, a pozrieme sa na jeho susediace vrcholy (vrcholy, s ktorými je prepojený hranou). Vyberieme náhodne jeden z nich (v príklade bod w), okrem vrcholu, s ktorým ho spája prvotne vybraná hrana (v príklade bod u). Pozrieme sa, či existuje hrana medzi týmito dvoma vrcholmi (u a w). Ak nie, s určitou pravdepodobnosťou ju doplníme. Túto pravdepodobnosť označujeme *pdt* (pravdepodobnosť doplnenia trojuholníka) a nastavujeme na začiatku ako jeden z parametrov modelu. Ďalším parametrom je M - počet opakovaní procesu dopĺňania na trojuholník. Tento parameter nastavujeme podľa počtu vrcholov grafu N na približne rovnakú hodnotu.



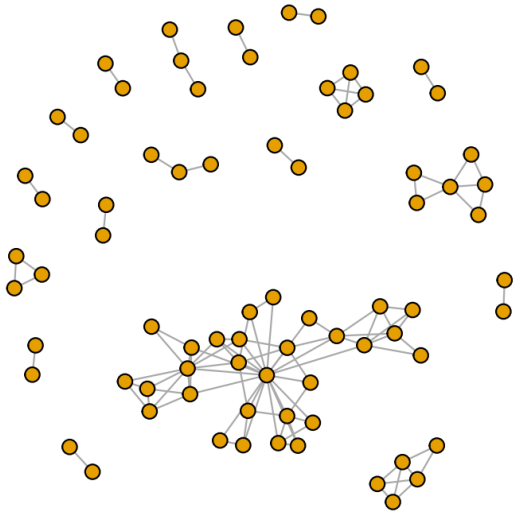
Obr. 15: Proces doplnenia hrany (kontaktu) do trojuholníka [4, str. 183].

Procesom dopĺňania trojuholníkov dosiahneme zhustenie jednotlivých komponentov v grafe, čím vytvoríme akési skupinky ľudí, ktoré sú navzájom v kontakte, ako napríklad príslučníci jednej domácnosti, školy, či športového klubu. V grafe pribudnú nové hrany, čím sa zvýšia aj stupne vrcholov, avšak ich postupnosť aj naďalej bude replikovať mocninové

rozdelenie. Je dôležité si uvedomiť, že aj keď proces opakujeme M-krát, neznamená to, že do grafu doplníme M hrán. Môže nastať prípad, kde hrana medzi dvoma vybranými vrcholmi už existuje (v príklade na obr. 15 by už existovala hrana medzi bodmi u a w). Ďalšou situáciou, kedy môže proces zlyhať, je prípad, keď vybraný vrchol zo zvolenej hrany má stupeň 1 a nemá žiadnych ďalších susedov, okrem vrcholu, s ktorým ho spája zvolená hrana. Tiež si treba uvedomiť, že pracujeme s pravdepodobnosťou doplnenia trojuholníka, ktorú môžeme nastaviť na hodnotu nižšiu ako 1.



Obr. 16: Graf s $N = 75$ vrcholmi pred doplnením trojuholníkov (vlastné spracovanie v [6]).



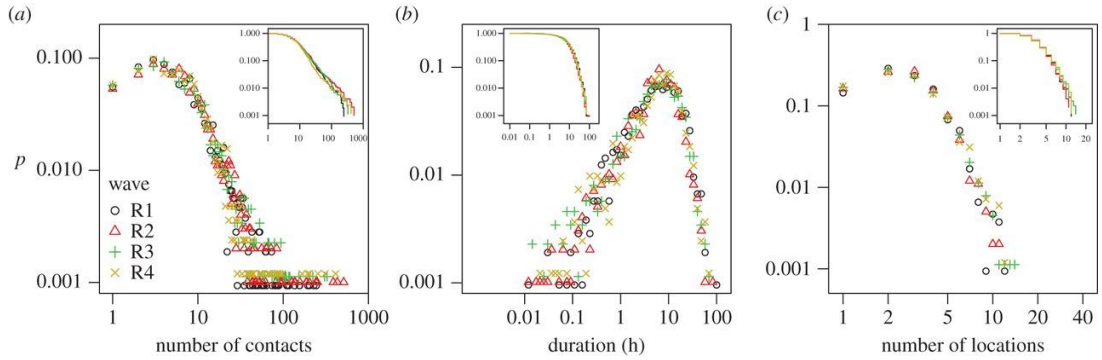
Obr. 17: Graf s $N = 75$ vrcholmi po doplnení trojuholníkov (vlastné spracovanie v [6]).

Na obr. 16 vidíme ilustračný príklad grafu s malým počtom $N = 75$ vrcholov, v ktorom sa nachádza jeden väčší a niekoľko menších komponentov. Hustota hrán v jednotlivých komponentoch je nízka. Procesom dopĺňania hrán dostane graf na obr. 17, ktorý má tie isté komponenty, ale hustota hrán v nich sa zvýšila. Teda ľudia v jednotlivých skupinkách sú navzájom viac poprepájaní.

Dostávame sieť, ktorá sa už dostatočne približuje realite. Takýto jednoduchý graf nám hovorí, ktorí ľudia sú navzájom v kontakte, no nie každé naše spojenie má rovnakú intenzitu. Je oveľa väčšia pravdepodobnosť, že sa nakazíme od niekoho, s kým žijeme v jednej domácnosti alebo pracujeme v jednej kancelárii, ako že sa infikujeme od osoby, ktorú stretneme v MHD, v obchode alebo sa len chvíľu rozprávame na ulici. Chceli by sme nejako ohodnotiť jednotlivé hrany, čím dostaneme z nášho jednoduchého grafu vážený.

Ohodnocovanie hrán

Procesom ohodnocovania hrán zmeníme našu jednoduchú maticu susedností s hodnotami 0 a 1 na váženú maticu, ktorej zložky sa pohybujú v intervale $[0, 1]$. Vhodnou distribúciou vieme náhodne z požadovaného intervalu tieto hodnoty nastaviť. Túto distribúciu sme hľadali pomocou štúdie obyvateľov Hong Kongu, zameranej na časovú variáciu ľudských stretnutí [3].



Obr. 18: Normalizované rozdelenia (a) počtu kontaktov a (b) celkového trvania uskutočnených kontaktov a (c) počtu miest, v ktorých došlo ku kontaktom pre každú zo štyroch vĺn vzorkovania [3].

Na obrázku 18 b) môžeme vidieť pravdepodobnostnú distribúciu času trvania kontaktu pre jednotlivé kontakty. Urobíme predpoklad, že čas kontaktu je proporcionálny pravdepodobnosti infikovania (ak je jedna z osôb infikovaná). Zároveň však dlhé kontakty, nad 10 hodín, považujeme za veľmi blízke a pri nich definujeme pravdepodobnosť prenosu ochorenia ako 1.

Keďže závislosť na obrázku 18 b) sa javí na log-log škále v rozsahu od 0,01 hod. do 10 hod. ako lineárna, pre váhy hrán použijeme mocninovú distribúciu. Je potrebné však všetko príslušne normalizovať, keďže blízke kontakty nad 10 hodín už majú pravdepodobnosť prenosu 1.

Závislosť pravdepodobnosti p na obrázku 18 b) od trvania kontaktu (duration - d) na intervale 0,01 hod. až 10 hod. možno aproximovať závislosťou:

$$(\log p) + 1 = \frac{2}{3}\left((\log d) - 1\right), \tag{2}$$

pretože ide o lineárnu závislosť $\log p$ od $\log d$ prechádzajúcu bodmi $(10^{-2}, 10^{-3})$ a $(10^{1}, 10^{-1})$.

Ak túto závislosť exponenciujeme nad číslom 10, dostávame:

$$\begin{aligned} 10^{(logp)+1} &= (10^{(logd)-1})^{\frac{2}{3}} \\ 10p &= \left(\frac{d}{10}\right)^{\frac{2}{3}} \\ p &= \frac{1}{10}\left(\frac{d}{10}\right)^{\frac{2}{3}}. \end{aligned}$$

Táto pravdepodobnosť však nepokrýva všetky kontakty, ale len ich časť pod 10 hodín. Akú časť tvoria presne takéto kontakty zistíme integrovaním ako:

$$\int_0^{10} p(d)dd.$$

Integrujeme pre jednoduchosť od 0 a nie od 10^{-2}, lebo je to zanedbateľný rozdiel.

Vypočítame integrál:

$$\int_0^{10} 10^{-\frac{5}{3}} d^{\frac{2}{3}} dd = 10^{-\frac{5}{3}} \left[\frac{3}{5} d^{\frac{5}{3}}\right]_0^{10} = \frac{3}{5}.$$

Zároveň kumulatívna pravdepodobnosť dĺžky kontaktu aspoň d je daná integrálom:

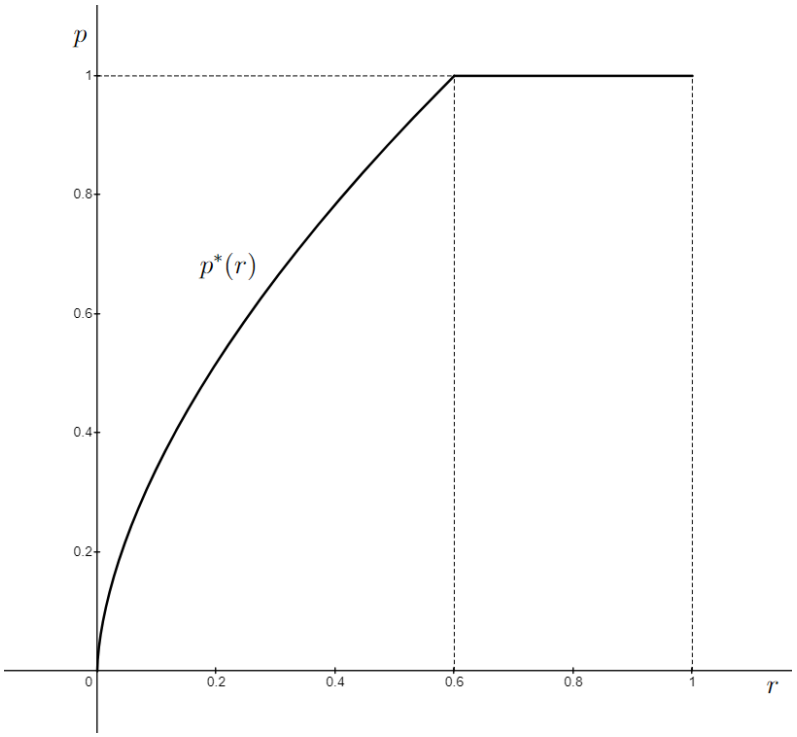
$$\int_0^{d} 10^{-\frac{5}{3}} s^{\frac{2}{3}} ds = \frac{3}{5}\left(\frac{d}{10}\right)^{\frac{5}{3}}.$$

Exponent závislosti kumulatívnej pravdepodobnosti dĺžky kontaktu pri kratších kontaktoch do 10 hodín je teda $\frac{5}{3}$.

Preto definujeme priraďovaciu funkciu $p^*(r)$:

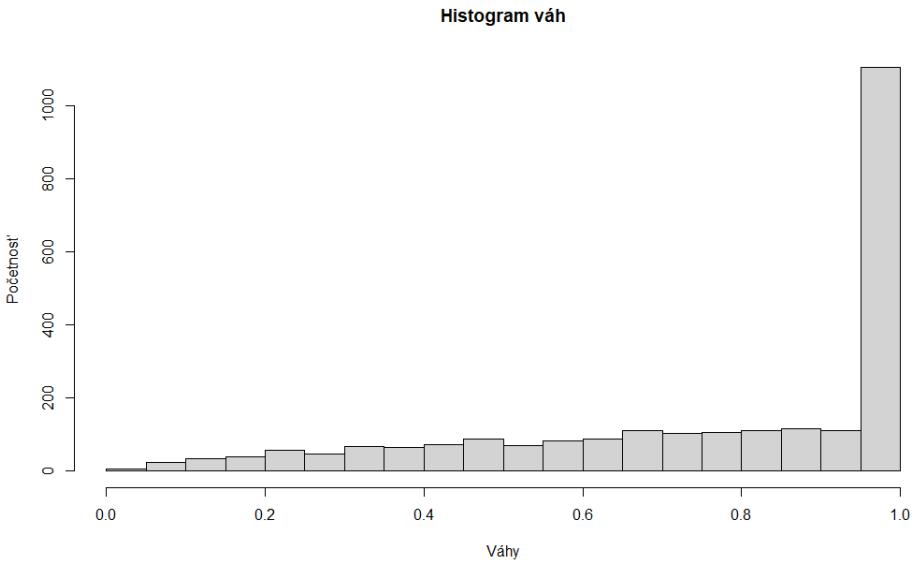
$$p^*(r) = \begin{cases} \frac{5}{3}\frac{3}{5} r^{\frac{3}{5}}, & \text{ak} \quad r \in \left[0, \frac{3}{5}\right], \\ 1, & \text{ak} \quad r \in \left(\frac{3}{5}, 1\right], \end{cases} \tag{3}$$

ktorá priradí náhodnému číslu r z intervalu $[0, 1]$ pravdepodobnosť $p^* = p^*(r)$, ktorú priradíme ako váhu (pravdepodobnosť prenosu infekcie cez túto hranu) náhodne vybranej hrane grafu. Táto funkcia priraďuje pravdepodobnosť 1 pre náhodné čísla vačšie ako $\frac{3}{5}$, t. j. blízke kontakty nad 10 hodín. Pre kratšie trvajúce kontakty je závislosť pravdepodobnosti mocninová s exponentom $\frac{5}{3}$ odvodeným z mocninovej závislosti dát, taká, že pre $r = 0$ priraďuje nulovú pravdepodobonosť $p^* = 0$ prenosu. Tá však mocninovo rastie, až sa spojito napojí v $r = \frac{3}{5}$ na hodnotu $p^* = 1$ (obr. 19).



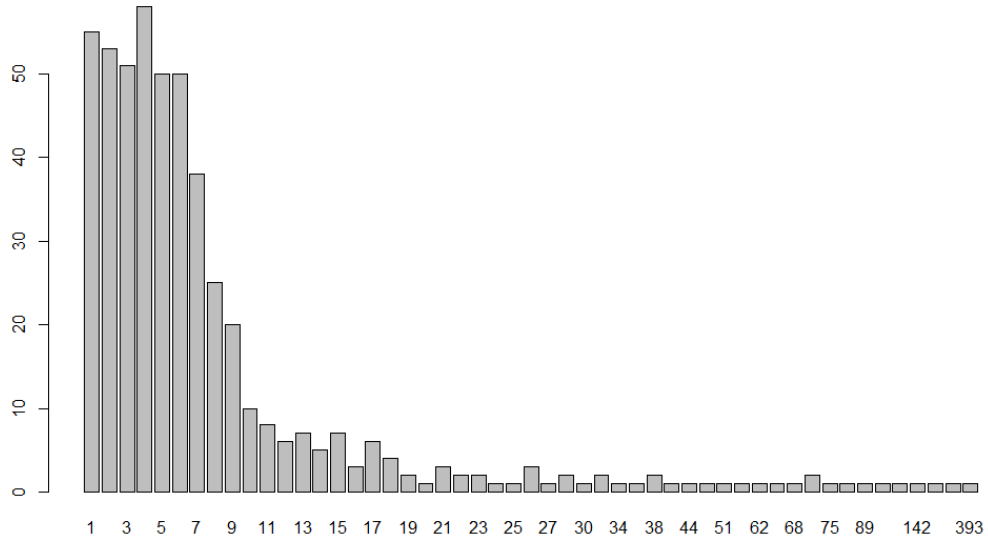
Obr. 19: Priraďovacia funkcia $p^*(r)$ (vlastné spracovanie).

Na obrázku 20 môžeme vidieť konkrétnu realizáciu početnosti hrán s danou váhou po použití priraďovacej funkcie $p^*(r)$ na hrany jednoduchého grafu (pre každú hranu sa náhodne vygeneruje číslo r a následne sa jej podľa funkcie pridelí váha).



Obr. 20: Početnosť hrán v grafe s danou váhou (vlastné spracovanie v [6]).

Na obrázku 18 a) môžeme vidieť pravdepodobnostnú distribúciu počtu kontaktov. Počet kontaktov sme modelovali pomocou mocninovej distribúcie s parametrom α. To sedí aj s daným grafom pre počet kontaktov vyšší ako 6. Preto sme sa rozhodli na základe tohto upraviť distribúciu, pomocou ktorej vytvárame postupnosť prvkov. Dieliky našej pomyselnej mriežky budú mať pre prvých 6 hodnôt rovnakú šírku, t. j. dieliky budú veľkosti $\frac{1}{6^\alpha}, \frac{1}{6^\alpha}, \frac{1}{6^\alpha}, \frac{1}{6^\alpha}, \frac{1}{6^\alpha}, \frac{1}{6^\alpha}, \frac{1}{7^\alpha}, \ldots, \frac{1}{N^\alpha}$, čím sa upravia aj jednotlivé pravdepodobnosti. Príklad postupnosti stupňov vrcholov pre túto upravenú distribúciu môžeme vidieť na obrázku 21.



Obr. 21: Histogram stupňov vrcholov pre upravenú distribúciu (vlastné spracovanie v [6]).

4.2 Simulácia epidémie

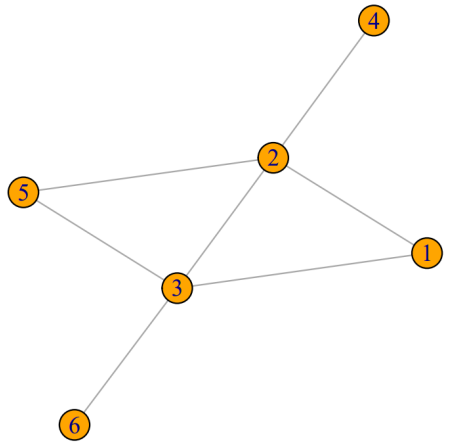
V predchádzajúcej kapitole sme popísali, ako vytvoriť graf reprezentujúci sieť ľudských kontaktov pomocou určitých parametrov, ktorý dostatočne presne predstavuje realitu. Na tomto grafe by sme chceli simulovať epidémiu a sledovať jej priebeh počas každého kroku. Na začiatku potrebujeme šírenie infekcie naštartovať. Náhodne spomedzi všetkých vrcholov vyberieme jeden, ktorému zmeníme stav na *infikovaný*. V druhom kroku by sme sa chceli pozrieť, koho môže tento infikovaný človek nakaziť, na čo využijeme maticu

susedností nášho grafu, ktorá nám hovorí o prepojení vrcholov grafu. Tiež si vytvoríme vektor novonakazených (nazveme ho *nov*), ktorý bude dĺžky N a bude obsahovať hodnoty 1 (ak je daný vrchol novonakazený) a 0 (ak nie je). Po prvom kroku dostaneme vektor so samými nulami, jediná jednotka bude v zložke reprezentujúcej prvotne nakazený bod.

Uvažujme najprv jednoduchý graf, ktorého všetky hrany majú rovnakú váhu, teda máme maticu susedností, ktorá obsahuje iba jednotky a nuly. V tomto prípade je pravdepodobnosť, že sa daný kontakt nakazí, rovná 1. Po prenásobení vektora novonakazených maticou susedností dostaneme nový vektor, ktorý bude obsahovať jednotky na zložkách reprezentujúcich susedov nakazeného vrcholu. V ďalších krokoch budeme tiež prenásobovať nový vektor novonakazených upravenou maticou susedností, avšak pribudnú nám ďalšie potrebné úpravy vektorov a matíc.

Podrobne si to vysvetlíme na nasledujúcom ilustračnom príklade. Máme graf s $N = 6$ vrcholmi, znázornený na obr. 22. Matica susedností A_1 tohto grafu bude nasledovná:

$$A_1 = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}.$$



Obr. 22: Graf s $N = 6$ vrcholmi pred šírením epidémie (vlastné spracovanie v [6]).

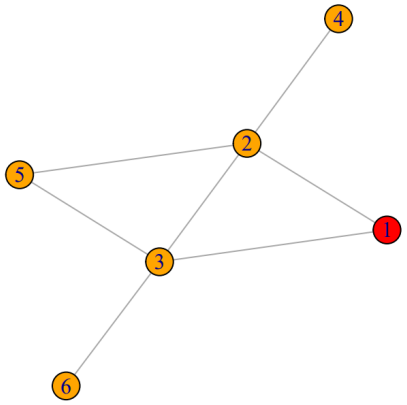
Náhodne nakazíme prvého človeka, v našom príklade to bude vrchol 1 (obr. 23). Vektor novonakazených v prvom kroku bude $nov_1 = (1, 0, 0, 0, 0, 0)$. Pred násobením v každom kroku odstránime v matici susedností stĺpce s indexmi vrcholov, ktorí sú novonakazení, aby sme pri vyššom počte novonakazených zabránili opätovnému nakazeniu už infikovaného vrcholu (dôležitosť tohto odstránenia sa ukáže v druhom kroku tohto príkladu).

Teraz vynásobíme daný vektor novonakazených upravenou maticou susedností $\tilde{A}_1$ a dostaneme nový vektor nakazených:

$$nov_1 \cdot \tilde{A}_1 = (1,0,0,0,0,0) \cdot \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} = (0,1,1,0,0,0) = nov_2.$$

Dostali sme nový vektor novonakazených nov_2, ktorý nám hovorí, že v prvom kroku sa infikovali vrcholy 2 a 3 (obr. 24), pričom vrchol 1 už neuvažujeme. Po prenásobení odstránime riadky a stĺpce matice susedností prislúchajúce vrcholom, ktoré boli na začiatku tohto kroku infikované. Tým tieto vrcholy už nebudeme ďalej v priebehu epidémie uvažovať. Dostávame maticu susedností A_2:

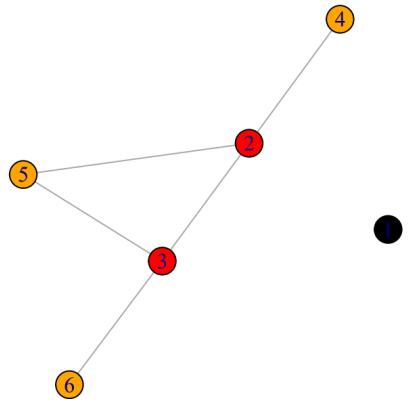
$$A_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}.$$



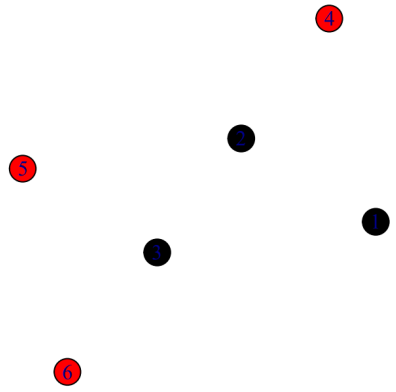
Obr. 23: Graf v prvom kroku s prvým infikovaným vrcholom (vlastné spracovanie v [6]).

Opäť pred prenásobením odstránime stĺpce vrcholov, ktoré sú práve nakazené a dostaneme upravenú maticu susedností $\tilde{A}_2$. Druhý krok bude:

$$nov_2 \cdot \tilde{A}_2 = (0,1,1,0,0,0) \cdot \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} = (0,0,0,1,2,1) = nov_3.$$



Obr. 24: Graf v druhom kroku s dvoma novými infikovanými vrcholmi (vlastné spracovanie v [6]).



Obr. 25: Graf v treťom kroku, všetky vrcholy sú už infikované (vlastné spracovanie v [6]).

Ak by sme maticu neupravili a použili A_2, dostali by sme vektor $nov_3 = (0,1,1,1,2,1)$, ktorý by nám hovoril, že vrcholy 2 a 3 sa opäť nakazili, čo ale nedáva zmysel. Preto je dôležité stĺpce odstraňovať. Vo vektore novonakazených po tomto kroku vidíme aj hodnoty väčšie ako 1, čo je spôsobené tým, že daný bod sa mohol nakaziť od viacerých infikovaných. Tieto hodnoty väčšie ako 1 iba zmeníme na 1, čím dostaneme vektor $nov_3 = (0,0,0,1,1,1)$. Ten nám hovorí, že v tomto kroku sa nakazili vrcholy 4, 5 a 6 (obr. 25), vrcholy 1, 2 a 3 už neuvažujeme. Na obrázku vidíme, že všetky vrcholy sú nakazené, čím vlna tejto epidémie skončila. V programe by sme dostali vektor novonakazených zložený zo samých núl, čím program ukončíme.

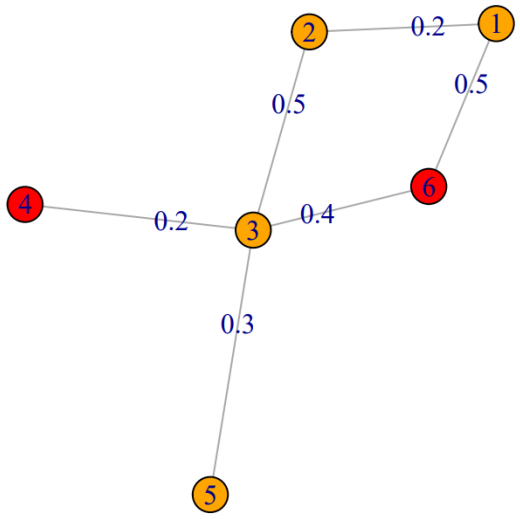
Toto bol nenáročný príklad priebehu epidémie pre jednoduchý graf. My však budeme používať graf vážený, ktorý sa viac približuje realite. Ako vieme, matica susedností takéhoto grafu neobsahuje iba nuly a jednotky, ale aj hodnoty medzi nimi, podľa váhy hrany spájajúcej príslušné dva vrcholy. V tomto prípade už nemôžeme používať jednoduché prenásobovanie vektora novonakazených s maticou susedností, ale musíme predpokladať, že človek (vrchol) sa môže infikovať od nakazeného (ak je s ním v kontakte) s pravdepodobnosťou, ktorá je reprezentovaná váhou hrany, ktorá tieto dva vrcholy spája.

Zavedieme si vzorec na výpočet pravdepodobnosti nakazenia i-teho vrcholu:

$$p_i = 1 - \prod_{j \in J} (1 - a_{ij}), \tag{4}$$

kde množina J bude množinou tých indexov vrcholov, ktoré sú práve nakazené (vo vektore novonakazených majú hodnotu 1). Ukážeme si to na konkrétnom príklade jedného kroku epidémie. Uvažujme výsek 6 vrcholov grafu (obr. 26) s príslušnou maticou susedností A:

$$A = \begin{pmatrix} 0 & 0,2 & 0 & 0 & 0 & 0,5 \\ 0,2 & 0 & 0,5 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0,2 & 0,3 & 0,4 \\ 0 & 0 & 0,2 & 0 & 0 & 0 \\ 0 & 0 & 0,3 & 0 & 0 & 0 \\ 0.5 & 0 & 0,4 & 0 & 0 & 0 \end{pmatrix}.$$



Obr. 26: Výsek 6 vrcholov váženého grafu (vlastné spracovanie v [6]).

Prepokladajme, že na začiatku tohto kroku máme vektor novonakazených $nov = (0, 0, 0, 1, 0, 1)$, teda infikované sú vrcholy 4 a 6. V našom programe začneme tým, že si vytvoríme pomocnú maticu C, v ktorej vynulujeme riadky prislúchajúce vrcholom, ktoré sú infikované, keďže nechceme, aby sa tieto vrcholy mohli opätovne nakaziť. Vo vzorci (4) pre výpočet pravdepodobnosti nakazenia vrcholu i v produkte počítame iba

s vrcholmi j, ktoré sú infikované. V programe to vyriešime vynulovaním stĺpcov, ktoré reprezentujú nenakazené vrcholy. Bolo by náročné nastaviť produkt tak, aby počítal iba s danými indexami, a tak využijeme doplnkovú pravdepodobnosť $1 - a_{ij}$, že vrchol j neinfikuje vrchol i. Teraz, keď spravíme produkt cez tieto doplnkové pravdepodobnosti, dostaneme celkovú pravdepodobnosť, že sa daný vrchol v tomto kroku neinfikuje. Ak by sme spravili produkt iba cez pravdepodobnosti a_{ij}, nulovosť niektorých prvkov by spôsobila, že by aj výsledná pravdepodobnosť bola nulová. Je nesprávne uvažovať, že celková pravdepodobnosť nakazenia je súčinom jednotlivých pravdepodobností, aj keby sa v tomto súčine žiadne nuly nenachádzali. Máme vypočítanú pravdepodobnosť, že sa daný vrchol neinfikuje, a k tomu doplnková pravdepodobnosť je práve p_i, že sa daný vrchol v tomto kroku infikuje. V našom ilustračnom príklade (obr. 26) teda dostávame pravdepodobnosti: $p_1 = 0,5, \quad p_2 = 0, \quad p_3 = 0,52, \quad p_4 = 0, \quad p_5 = 0, \quad p_6 = 0$. Týmto dostávame pravdepodobnostný model, ktorý je komplikovanejší ako jednoduchý, no oveľa viac sa približuje realite. To znamená, že aj keď sme boli v kontakte s niekým nakazeným, neznamená to, že sme sa s pravdepodobnosťou 100% od neho nakazili.

5 Vyhodnocovanie modelu

Predmetom záverečnej kapitoly bude zisťovať štatistiky vychádzajúce z modelu predstavenom v predošlej kapitole. Budeme sledovať viaceré charakteristiky grafu pri rôznych nastaveniach parametrov.

Pre vygenerovanú postupnosť stupňov N vrcholov budeme generovať G rôznych vážených grafov, podľa postupu vysvetľovanom v predošlej kapitole. Pre každý graf budeme simulovať Z epidémií, pričom zakaždým budeme generovať nový prvý náhodný vrchol, ktorým začneme nákazu. Teda celkovo jedno pozorovanie pre dané nastavenie parametrov a vygenerovanú postupnosť stupňov vrcholov bude pozostávať z $G \cdot Z$ cyklov. Pre každý cyklus budeme sledovať počet novonakazených v jednotlivých krokoch vlny. Taktiež budeme sledovať v každom kroku maximálnu vlastnú hodnotu, resp. maximálnu absolútnu hodnotu vlastnej hodnoty (spektrálny polomer) aktuálnej matice susedností (matica susedností s nulovými riadkami a stĺpcami pre už nakazené vrcholy). Každým krokom bude táto matica redšia. Z každého cyklu dostaneme vektor počtu novonakazených a maximálnych vlastných hodnôt matice susedností. Tieto vektory budú mať rovnakú dĺžku, ktorá reprezentuje trvanie jednej vlny epidémie pre daný vygenerovaný graf. Celkovo dostaneme zoznam $G \cdot Z$ takýchto vektorov. Z nich spravíme priemer pre jednotlivé kroky (dni) a získame výsledný vektor. Vektor novonakazených nám hovorí o tom, koľko ľudí sa v daný deň infikovalo, označíme ho n. Nás by zaujímal aj aktuálny počet nakazených, teda aj ľudí, ktorí sa nakazili v predošlé dni. Spravíme konvolúciu počtu novonakazených z vektora n. Zavedieme si nový vektor ako parameter, filter h. Zložky tohto vektora nám hovoria o tom, koľko ľudí z daného dňa je ešte nakazených. Prvá zložka hovorí, aký podiel z dnes nakazených je infikovaných, druhá zložka aký pomer ľudí, ktorí sa včera nakazili, je dnes infikovaných atď. Na výpočet aktuálneho stavu nakazených v deň d použijeme vzorec:

$$s_d = \sum_{i=1}^{d} n_{d-i+1} \cdot h_i. \tag{5}$$

Pomocou tohto vzorca dostaneme aktuálny stav nakazených, avšak v reálnom svete nepoznáme tieto počty. Vieme iba o tých, ktorí sa nejakým spôsobom otestovali a zistili, že sú infikovaní. Zavedieme nový vektor náhodného testovania t, ktorý nám hovorí, s akou pravdepodobnosťou sa ľudia v daný deň otestovali a zistili svoju infekciu. Na výpočet

počtu zistených infikovaných v deň d použijeme vzorec:

$$z_d = \sum_{i=1}^{d} n_{d-i+1} \cdot h_i \cdot t_i. \tag{6}$$

Dostávame tri vektory, ktoré nám hovoria o počtoch novonakazených, aktuálne nakazených a zistených nakazených. Pozrieme sa na zmeny týchto počtov. Vytvoríme nové vektory:

$$\begin{aligned}
\tilde{n} &= \frac{n_{i+1}}{n_i}, \quad i = 1, \ldots, l-1, \\
\tilde{s} &= \frac{s_{i+1}}{s_i}, \quad i = 1, \ldots, l-1, \\
\tilde{z} &= \frac{z_{i+1}}{z_i}, \quad i = 1, \ldots, l-1,
\end{aligned}$$

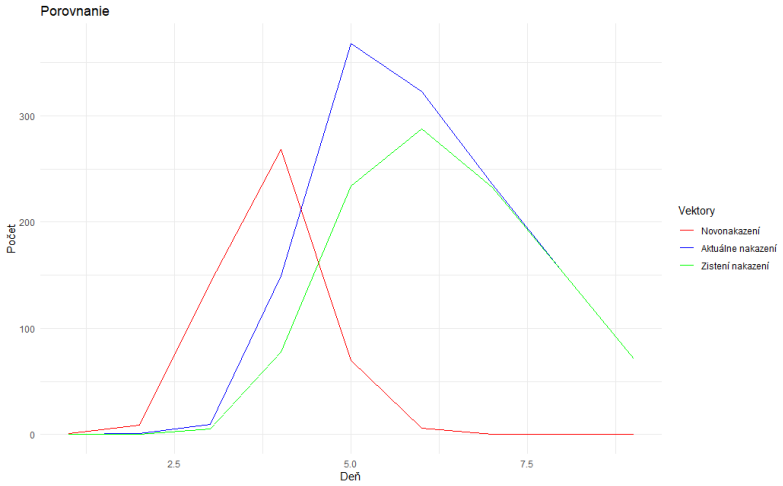
kde l je dĺžka vektora novonakazených. Práve tieto zmeny počtu nakazených majú niečo spoločné s hodnotou reprodukčného čísla, o ktorom sme hovorili v prvej kapitole.

5.1 Realizácie pozorovaní

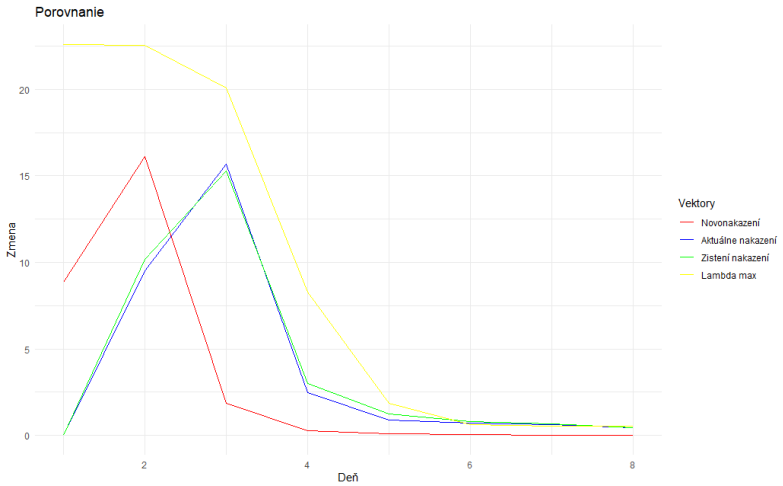
Pozrieme sa na pár ilustračných príkladov s konkrétnym nastavením parametrov vstupujúcich do tohto modelu. Tými sú počet vrcholov N, parameter mocninového rozdelenia stupňov α, počet opakovaní dopĺňania na trojuholník M, pravdepodobnosť doplnenia na trojuholník pdt, pravdepodobnosť nakazenia pn, počet generovania grafov G, počet simulácií pre jednotlivé grafy Z a filtre h a t. Keďže pracujeme s rôznymi pravdepodobnosťami, môžeme hovoriť o našom modeli ako o pravdepodobnostnom modeli.

Prvým nastavením parametrov bude:
$N = 500$, $\alpha = 2.4$, $M = 500$, $pdt = 0.9$, $pn = 1$, $G = 100$, $Z = 10$,
$h = (0, 1, \frac{2}{3}, \frac{1}{2}, \frac{1}{3}, \frac{1}{6}, 0, 0, 0, \ldots)$, $\quad t = (\frac{1}{2}, \frac{1}{2}, 1, 1, 1, 1, \frac{1}{2}, \frac{1}{2}, \frac{1}{4}, \frac{1}{4}, 0, 0, 0, \ldots)$.

Pre dané nastavenie parametrov dostávame graf pre počty nakazených na obr. 27. Nás však bude zaujímať graf zmien týchto počtov spojený s krivkou maximálnych vlastných hodnôt (obr. 28). Z neho môžeme vidieť, že krivky zmien aktuálneho a zisteného stavu sú približne rovnaké. Krivka zmien počtu novonakazených predbieha tieto dve krivky, resp. je medzi nimi určitý posun. Tiež si môžeme všimnúť, že krivka maximálnych vlastných hodnôt detekuje zlomy krivky zmien novonakazených.



Obr. 27: Počty nakazených v jednej vlne epidémie pre dané nastavenie parametrov (vlastné spracovanie v [6]).



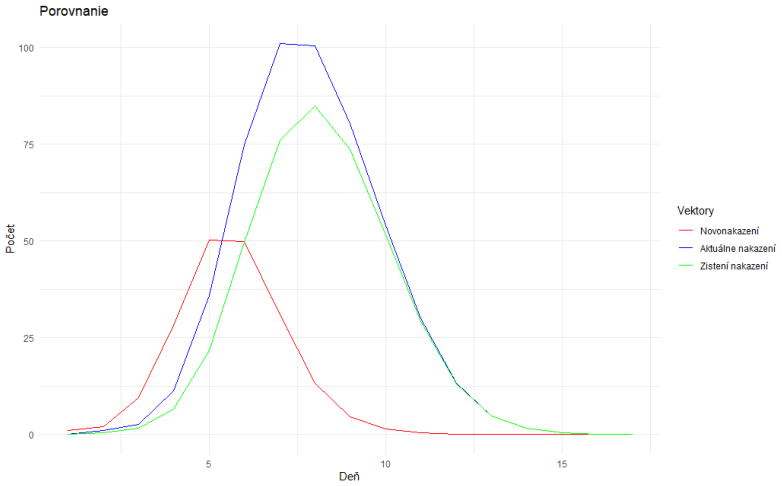
Obr. 28: Porovnanie kriviek $\tilde{n}, \tilde{s}, \tilde{z}$ a vektora maximálnej vlastnej hodnoty (vlastné spracovanie v [6]).

V druhom prípade zmeníme pravdepodobnosť nakazenia pn na menšiu hodnotu, čím predpokladáme nižší priemerný počet nakazených v jednotlivé dni, a taktiež predĺženie trvania jednej vlny epidémie. Nastavenie parametrov bude nasledovné:

$N = 500$, $\alpha = 2.4$, $M = 500$, $pdt = 0.9$, $pn = 0.25$, $G = 100$, $Z = 10$,

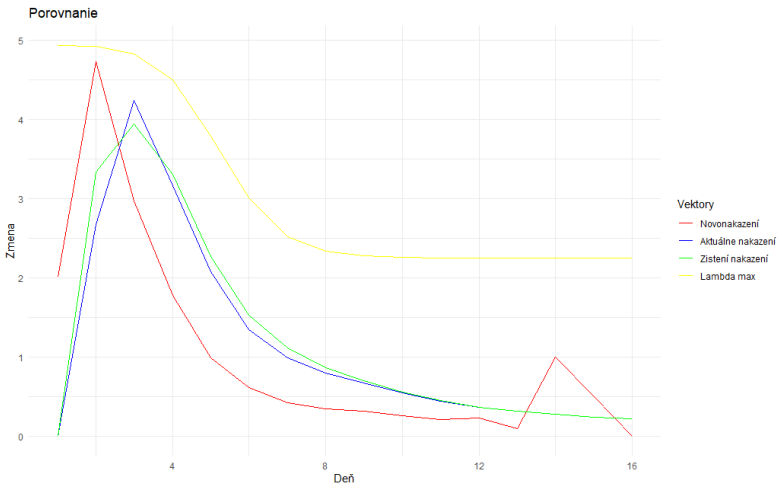
$h = (0, 1, \frac{2}{3}, \frac{1}{2}, \frac{1}{3}, \frac{1}{6}, 0, 0, 0, \dots)$, $\quad t = (\frac{1}{2}, \frac{1}{2}, 1, 1, 1, 1, \frac{1}{2}, \frac{1}{2}, \frac{1}{4}, \frac{1}{4}, 0, 0, 0, \dots)$.

Graf počtov nakazených vidíme na obr. 29. Ako sme predpokladali, dĺžka trvania vlny sa zvýšila z 9 na 17 dní.



Obr. 29: Počty nakazených v jednej vlne epidémie pre nižšiu pravdepodobnosť nakazenia (vlastné spracovanie v [6]).

Pozrieme sa opäť aj na graf zmien (obr. 30). Krivky zmien aktuálneho a zisteného stavu nakazených sú znovu približne rovnaké. Krivka zmien počtu novonakazených je oproti nim posunutá.



Obr. 30: Porovnanie kriviek $\tilde{n}, \tilde{s}, \tilde{z}$ a vektora maximálnej vlastnej hodnoty pre znížený pravdepodobnosť nakazenia (vlastné spracovanie v [6]).

5.2 Analýza grafov

Analýzou získaných grafov by sme chceli zistiť a presnejšie popísať vzťah medzi jednotlivými krivkami zmien počtu nakazených a krivkou maximálnych vlastných hodnôt.

5.2.1 Vzťah kriviek zmien počtu nakazených

Najprv sa pozrieme na krivky zmien počtu novonakazených $\tilde{n}$, aktuálneho stavu nakazených $\tilde{s}$ a zistených nakazených $\tilde{z}$. Pri porovnávaní týchto kriviek budeme skúmať ich vzájomný posun a násobok. Na túto analýzu budeme využívať nasledujúcu minimalizačnú úlohu:

$$\min_{\alpha,\tau} \left\| \frac{\tilde{s}(t) - \alpha\tilde{n}(t-\tau)}{\tilde{s}(t)} \right\|, \tag{7}$$

kde hľadáme optimálny posun τ a násobok α. Rovnakú úlohu by sme riešili aj pre vektor zistených nakazených $\tilde{z}$, výmenou za vektor $\tilde{n}$, resp. $\tilde{s}$. Bez ujmy na všeobecnosti budeme ďalej pracovať s vektormi $\tilde{n}$ a $\tilde{s}$.

Vzhľadom na to, že máme diskrétne hodnoty v čase s dĺžkou kroku jeden deň, chceli by sme sa viac priblížiť k spojitému prípadu. Na to využijeme interpoláciu vektorov. Do našich vektorov zmien počtu nakazených doplníme hodnoty medzi jednotlivé body tak, že dĺžku kroku skrátime na štvrtinu dňa. Medzi každé dve zložky týchto vektorov doplníme práve tri hodnoty na základe lineárnej interpolácie (napr. na časovej osi medzi časy t_1 a t_2 doplníme časy $t_{1\frac{1}{4}}$, $t_{1\frac{2}{4}}$ a $t_{1\frac{3}{4}}$; nech zmena počtu nakazených v čase t_1 bola 1 a v čase t_2 9, tak pre doplnené časy budú interpolované hodnoty 3, 5 a 7). Vďaka tejto interpolácii dokážeme získať presnejšie výsledky minimalizačnej úlohy (7) s presnosťou na štvrtiny dňa.

Keďže naše vektory zmien počtu nakazených sú konečné vektory a optimalizujeme túto úlohu aj na základe posunu, pre $|\tau| > 0$ nemôžeme pracovať s celými vektormi. Ak máme vektory dĺžky k, tak pre jednotlivé τ berieme vektory dĺžky $k - |\tau|$. Pre $\tau > 0$, budeme brať z vektora $\tilde{s}$ zložky $\tau + 1, \ldots, k$ a z vektora $\tilde{n}$ zložky $1, \ldots, k - \tau$. Naopak, pre $\tau < 0$ z vektora $\tilde{s}$ beriem zložky $1, \ldots, k - |\tau|$ a z vekotra $\tilde{n}$ zložky $|\tau| + 1, \ldots, k$.

Ak by sme riešili minimalizačnú úlohu (7) s takto upravenými vektormi s rôznymi dĺžkami, pre veľké hodnoty posunu τ by sme pracovali s vektormi, ktorých dĺžka by bola malá, teda by sme porovnávali menej hodnôt. Preto našu minimalizačnú úlohu štandardizujeme na nasledovný tvar, ktorý zohladní dĺžky vektorov:

$$\min_{\alpha,\tau} \frac{1}{k - |\tau|} \left\| \frac{\tilde{s}(t) - \alpha\tilde{n}(t-\tau)}{\tilde{s}(t)} \right\|, \tag{8}$$

kde k je dĺžka už interpolovaného vektora.

Vzhľadom na to, že jednotlivé premenné vyjadrujeme jednu pomocou druhej, vieme tento problém vyjadriť aj matematicky. Hodnoty počtu aktuálneho stavu nakazených s v deň d sme si definovali pomocou počtu novonakazených n a filtra h ako:

$$s_d = \sum_{i=1}^{d} n_{d-i+1} \cdot h_i.$$

Napríklad počet aktuálne nakazených v piaty deň vypočítame ako $s_5 = n_5h_1 + n_4h_2 + \cdots + n_1h_5$. Tento počet vieme vyjadriť ekvivalentne v spojitom svete pomocou integrálu:

$$s(t) = \int_{-\infty}^{\infty} n(t-z)h(z)dz, \tag{9}$$

kde o filtri $h(z)$ vieme, že nadobúda hodnotu 0 pre z záporné a nezáporné hodnoty pre z nezáporné.

Predstavme si triviálny prípad, kedy by sme mali filter h definovaný pomocou delta funkcie nasledovne:

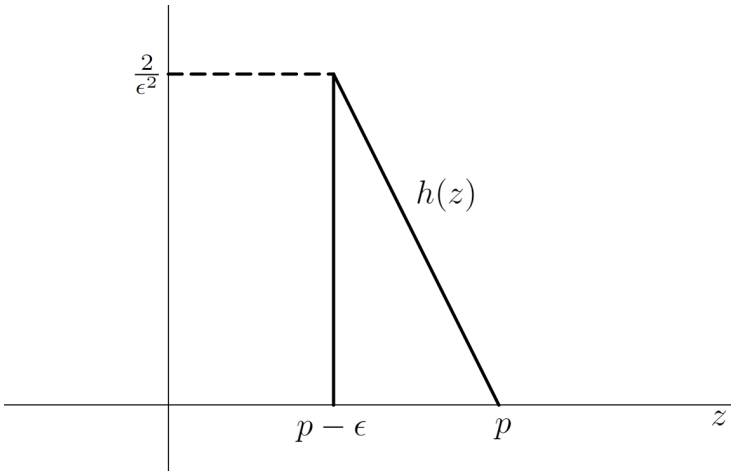
$$h_i = \begin{cases} 1, & i = p, \\ 0, & i \neq p, \end{cases} \quad \text{resp.} \quad h(z) = \begin{cases} \infty, & z = p, \\ 0, & z \neq p. \end{cases} \tag{10}$$

Pre takto definovaný filter by sme dostali $s_d = n_{d-p}$, resp. $s(t) = n(t-p)$. Teda pre lokalizovaný filter dostávame jednoznačné vyjadrenie počtu skutočne nakazených pomocou posunu p.

Uvažujme teraz zložitejší filter, ktorý definujeme ako "trojuholník"nasledovne (obr. 31):

$$h(z) = \begin{cases} 0, & z < p - \epsilon, \\ \frac{2}{\epsilon^2}(p - z), & p - \epsilon \leq z \leq p, \\ 0, & z > p, \end{cases} \tag{11}$$

pričom platí, že filter je normalizovaný, t. j. $\int_{-\infty}^{\infty} h(z)dz = 1$.



Obr. 31: Graf funkcie filtra $h(z)$ (vlastné spracovanie).

Pomocou konvolučnej vety a Fourierovej transformácie si upravíme (9) ako:

$$\hat{s}(\xi) = \hat{n}(\xi) \times \hat{h}(\xi). \tag{12}$$

Spočítajme Fourierovu transformáciu filtra $h(z)$:

$$\hat{h}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} h(z)e^{-i\xi z}dz = \frac{1}{\sqrt{2\pi}} \int_{p-\epsilon}^{p} \frac{2}{\epsilon^2}(p - z)e^{-i\xi z}dz.$$

Použijeme substitúciu $s = p - z$, čím dostaneme:

$$\frac{1}{\sqrt{2\pi}} \frac{2}{\epsilon^2} \int_{\epsilon}^{0} se^{-i\xi(p-s)}(-ds) = \frac{1}{\sqrt{2\pi}} \frac{2}{\epsilon^2} e^{-i\xi p} \int_{0}^{\epsilon} se^{i\xi s}ds.$$

Daný integrál vypočítame metódou per partes:

$$\begin{aligned}\int_0^\epsilon se^{i\xi s}ds &= \begin{bmatrix} u=s & v'=e^{i\xi s} \\ u'=1 & v=\dfrac{1}{i\xi}e^{i\xi s}\end{bmatrix} = \left[s\frac{1}{i\xi}e^{i\xi s}\right]_0^\epsilon - \int_0^\epsilon \frac{1}{i\xi}e^{i\xi s}ds \\ &= \frac{\epsilon}{i\xi}e^{i\xi\epsilon} - \frac{1}{i\xi}\left[\frac{1}{i\xi}e^{i\xi s}\right]_0^\epsilon = \frac{\epsilon}{i\xi}e^{i\xi\epsilon} + \frac{1}{\xi^2}\left(e^{i\xi\epsilon}-1\right).\end{aligned}$$

Teda Fourierova transformácia filtra je:

$$\hat{h}(\xi) = \frac{1}{\sqrt{2\pi}}\frac{2}{\epsilon^2}e^{-i\xi p}\left[\frac{\epsilon}{i\xi}e^{i\xi\epsilon} + \frac{1}{\xi^2}\left(e^{i\xi\epsilon}-1\right)\right]. \tag{13}$$

Pre $\epsilon \ll 1$ uvažujme nasledujúcu aproximáciu pomocou Taylorovho rozvoja:

$$e^x \approx 1 + x + \frac{x^2}{2}.$$

Pomocou tejto aproximácie si odhadneme aj hodnotu filtra:

$$\begin{aligned}\hat{h}(\xi) &\approx \frac{1}{\sqrt{2\pi}}\frac{2}{\epsilon^2}e^{-i\xi p}\left[\frac{\epsilon}{i\xi}\left(1+i\xi\epsilon+\frac{(i\xi\epsilon)^2}{2}\right)+\frac{1}{\xi^2}\left(1+i\xi\epsilon+\frac{(i\xi\epsilon)^2}{2}-1\right)\right] \\ &= \frac{1}{\sqrt{2\pi}}\frac{2}{\epsilon^2}e^{-i\xi p}\left[\frac{\epsilon}{i\xi}+\epsilon^2+\frac{i\epsilon^3\xi}{2}+\frac{i\epsilon}{\xi}-\frac{\epsilon^2}{2}\right] \\ &= \frac{1}{\sqrt{2\pi}}\frac{2}{\epsilon^2}e^{-i\xi p}\left(\frac{\epsilon^2}{2}+\frac{i\epsilon^3\xi}{2}\right), \\ \text{kde}\quad & \frac{\epsilon}{i\xi}+\frac{i\epsilon}{\xi} = \frac{\epsilon}{i\xi}+\frac{i^2\epsilon}{i\xi} = \frac{\epsilon}{i\xi}-\frac{\epsilon}{i\xi} = 0.\end{aligned}$$

Výraz môžeme ešte upraviť:

$$\frac{1}{\sqrt{2\pi}}\frac{2}{\epsilon^2}e^{-i\xi p}\left(\frac{\epsilon^2}{2}+\frac{i\epsilon^3\xi}{2}\right) = \frac{1}{\sqrt{2\pi}}\frac{2}{\epsilon^2}e^{-i\xi p}\left(\frac{\epsilon^2}{2}(1+i\epsilon\xi)\right) = \frac{1}{\sqrt{2\pi}}e^{-i\xi p}\left(1+i\epsilon\xi\right),$$

teda

$$\hat{h}(\xi) \approx \frac{1}{\sqrt{2\pi}}e^{-i\xi p}\left(1+i\epsilon\xi\right). \tag{14}$$

Teraz spätne môžeme vyjadriť počty aktuálne nakazených ako:

$$\begin{aligned}s(t) &= \hat{s}(\xi) = \hat{n}(\xi)\hat{h}(\xi) \approx \hat{n}(\xi)\frac{1}{\sqrt{2\pi}}e^{-i\xi p}\left(1+i\epsilon\xi\right) \\ &\approx n(t-p) + \epsilon\int_{-\infty}^{\infty}\ldots dz \end{aligned} \tag{15}$$

Ak uvažujeme $\epsilon \ll 1$, druhý člen výrazu (15) pôjde limitne do nuly, čiže môžeme aproximovať počet aktuálne nakazených pomocou počtu novonakazených s posunom p v čase:

$$s(t) \approx n(t-p). \tag{16}$$

Vo všeobecnosti pre ľubovolný filter h očakávame, že jeho Fourierova transformácia bude mať tvar:

$$\hat{h}(\xi) \approx \frac{1}{\sqrt{2\pi}} e^{-i\xi r} + \text{chyba}, \tag{17}$$

kde r je ťažisko filtra $h(z)$, teda platí:

$$\int_{-\infty}^{r} h(z)dz = \int_{r}^{\infty} h(z)dz. \tag{18}$$

Potom by sme vedeli vyjadriť vzťah počtu nakazených pomocou posunu r nasledovne:

$$s(t) \approx n(t-r) + \text{chyba}, \tag{19}$$

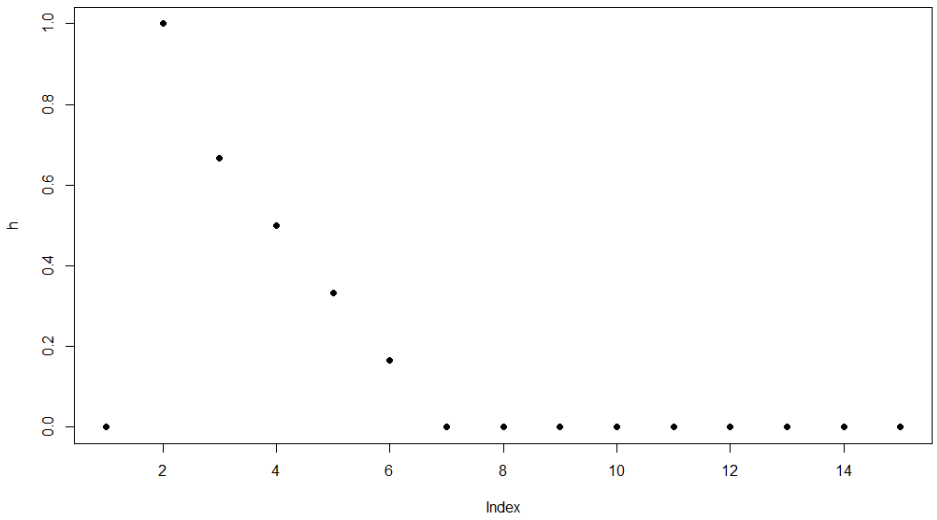
teda očakávame posun približne o ťažisko r.

Toto očakávanie by sme chceli overiť numericky pomocou nášho modelu a minimalizačnej úlohy (8). Ťažisko filtra r vypočítame podľa vzorca:

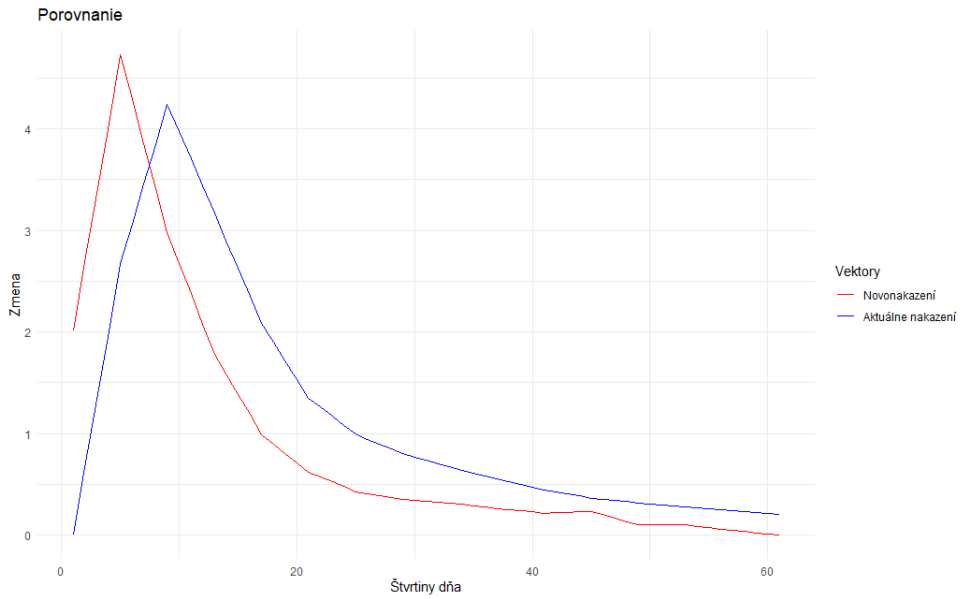
$$r = \frac{\sum_i i h_i}{\sum_i h_i}. \tag{20}$$

Na niekoľkých rôznych príkladoch nastavenia filtra h sa pozrieme na rozdiel medzi ťažiskom filtra r a optimálnym τ, ktoré bude výsledkom minimalizačnej úlohy.

Najprv uvažujme filter $h_1 = (0, 1, \frac{2}{3}, \frac{1}{2}, \frac{1}{3}, \frac{1}{6}, 0, 0, 0, \dots)$ zo vzorových príkladov v kapitole 5.1, ktorý môžeme vidieť vykreslený na obr. 32 . Ťažisko takéhoto filtra je $r = 3,25$. Pomocou optimalizačnej úlohy chceme zistiť optimálny posun medzi krivkami zmien počtu novonakazených $\tilde{n}$ a aktuálne nakazených $\tilde{s}$ (obr. 33). Vyšlo nám optimálne $\tilde{\tau} = 6$. Keďže sme vektory interpolovali, úloha rieši problém s presnosťou na štvrtiny dňa, teda $\tilde{\tau} = 6$ znamená optimálny posun 1,5 dňa.

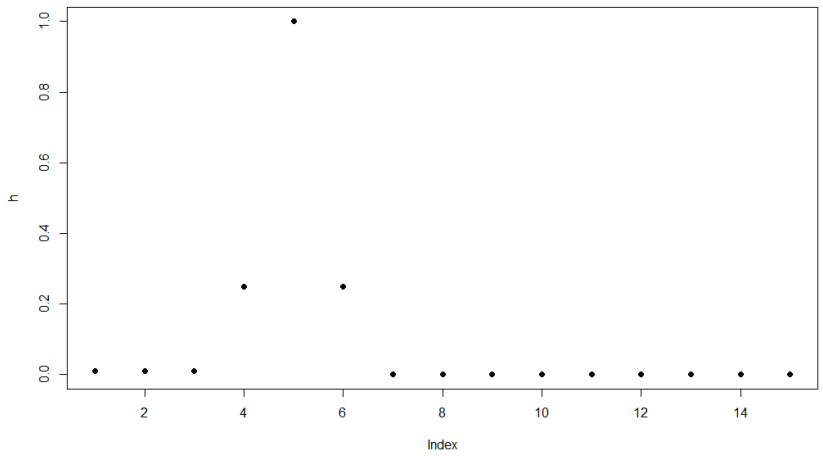


Obr. 32: Filter $h_1 = (0, 1, \frac{2}{3}, \frac{1}{2}, \frac{1}{3}, \frac{1}{6}, 0, 0, 0, \dots)$ (vlastné spracovanie v [6]).



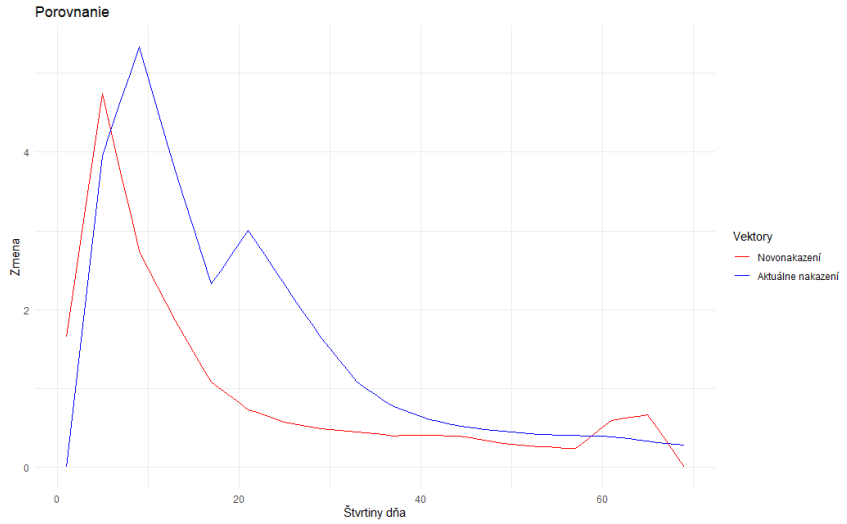
Obr. 33: Porovnanie interpolovaných kriviek zmien počtu novonakazených $\tilde{n}$ a aktuálne nakazených $\tilde{s}$ pre daný filter h_1 (vlastné spracovanie v [6]).

V prvom prípade nám vyšlo ťažisko filtra väčšie ako optimálny posun kriviek. Pozrime sa na druhý prípad, kde budeme uvažovať filter $h_2 = (0, 0, 0\frac{1}{4}, 1, \frac{1}{4}, 0, 0, 0, \dots)$, vykreslený na obr. 34. Ako je aj vidno z obrázku, keďže je filter symetrický, ťažisko je v bode $r = 5$.



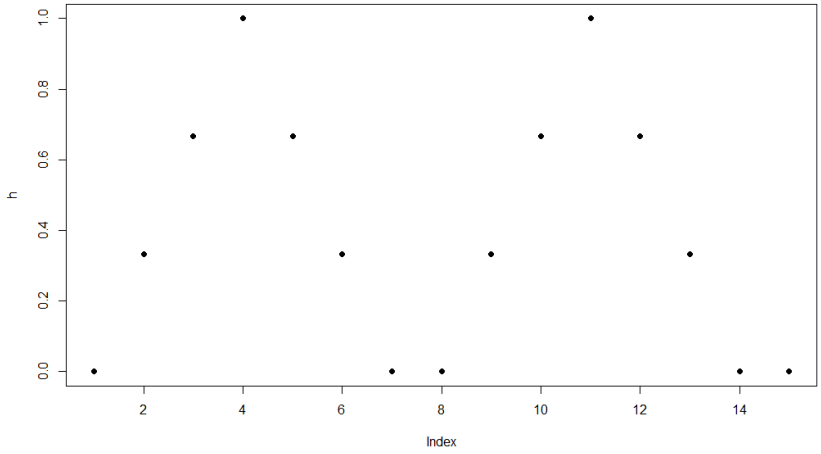
Obr. 34: Filter $h_2 = (0, 0, 0\frac{1}{4}, 1, \frac{1}{4}, 0, 0, 0, \dots)$ (vlastné spracovanie v [6]).

Opäť pomocou minimalizačnej úlohy chceme zistiť optimálny posun medzi krivkami $\tilde{n}$ a $\tilde{s}$ (obr. 33). Ten v tomto prípade vyšiel $\tilde{\tau} = 18$, čo je 4,5 dňa. Vidíme, že pre prípad lokalizovanejšieho filtra nám vyšiel optimálny posun približne rovný ťažisku.



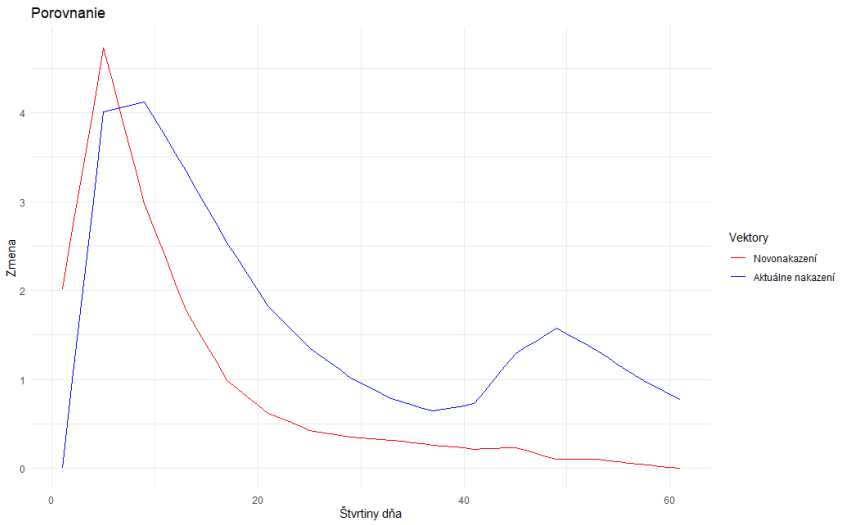
Obr. 35: Porovnanie interpolovaných kriviek zmien počtu novonakazených $\tilde{n}$ a aktuálne nakazených $\tilde{s}$ pre daný filter h_2 (vlastné spracovanie v [6]).

Doteraz sme sa pozerali na unimodálne lokalizované filtre h_1, h_2. Predstavme si, že by sme však mali filter $h_3 = (0, \frac{1}{3}, \frac{2}{3}, 1, \frac{2}{3}, \frac{1}{3}, 0, 0, \frac{1}{3}, \frac{2}{3}, 1, \frac{2}{3}, \frac{1}{3}, 0, 0, 0, \dots)$, vykreslený na obr. 36. Jeho ťažisko je v bode $r = 7,5$.



Obr. 36: Filter $h_3 = (0, \frac{1}{3}, \frac{2}{3}, 1, \frac{2}{3}, \frac{1}{3}, 0, 0, \frac{1}{3}, \frac{2}{3}, 1, \frac{2}{3}, \frac{1}{3}, 0, 0, 0, \dots)$ (vlastné spracovanie v [6]).

Vyriešením minimalizačnej úlohy pre krivky $\tilde{n}$ a $\tilde{s}$ (obr. 37) dostaneme optimálny posun $\tilde{\tau} = 8$, čo sú presne 2 dni. Naša teoretická predpoveď, že ťažisko filtra sa bude približne rovnať optimálnemu posunu sa nezhoduje s numerickými výsledkami pre ľubovoľný filter.



Obr. 37: Porovnanie interpolovaných kriviek zmien počtu novonakazených $\tilde{n}$ a aktuálne nakazených $\tilde{s}$ pre daný filter h_3 (vlastné spracovanie v [6]).

5.2.2 Vzťah zmeny počtu nakazených a maximálnej vlastnej hodnoty

Nech matica A_n je $N \times N$ matica susedností v čase t_n počas vlny pandémie. Nech $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ sú jej vlastné hodnoty a $w_1, w_2, \ldots, w_N$ sú jej ľavé ortogonálne vlastné vektory, pre ktoré platí:

$$\begin{aligned} \langle w_i, w_j \rangle &= 0, \quad i \neq j, \\ \langle w_i, w_j \rangle &= 1, \quad i = j. \end{aligned} \tag{21}$$

Najprv uvažujme neváženú maticu, teda maticu s hodnotami 0 a 1. Definujeme vektor novonakazených v čase t_n nasledovne:

$$v_i = \begin{cases} 1, & \text{ak je vrchol s indexom} \quad i \quad \text{práve infikovaný}, \\ 0, & \text{inak.} \end{cases} \tag{22}$$

Teda počet novoinfikovaných v čase t_n vieme spočítať ako $x_n = \sum_{i=1}^{N} v_i$. Po prenásobení vektora v maticou A_n dostávame nový vektor c, ktorého jednotlivé zložky c_i nám hovoria o počte susedov vrcholu i, ktorí sú infikovaní (v čase t_n), t. j. počet všetkých infikovaných kontaktov tohto vrcholu. Zapíšme si vektor novonakazených ako lineárnu kombináciu v báze vlastných vektorov w_i nasledovne:

$$v = \sum_{i=1}^{N} \alpha_i w_i. \tag{23}$$

Potom vďaka vlastnosti vlastných vektorov (21) vieme zapísať počet novoinfikovaných v čase t_n aj nasledovne:

$$x_n = ||v||^2 = \left\langle \sum_{i=1}^{N} \alpha_i w_i, \sum_{i=1}^{N} \alpha_i w_i \right\rangle = \sum_{i=1}^{N} \alpha_i^2. \tag{24}$$

Vektor c vieme potom zapísať ako:

$$c = vA_n = \sum_{i=1}^{N} \alpha_i w_i A_n = \sum_{i=1}^{N} \alpha_i \lambda_i w_i, \tag{25}$$

kde λ_i sú vlastné hodnoty matice A_n prislúchajúce vlastným vektorom w_i.

Opäť využitím vlastnosti (21) a (25) môžeme druhú mocninu normy vektoru c vyjadriť ako:

$$||c||^2 = \sum_{i=1}^{N} c_i^2 = \left\langle \sum_{i=1}^{N} \alpha_i \lambda_i w_i, \sum_{i=1}^{N} \alpha_i \lambda_i w_i \right\rangle = \sum_{i=1}^{N} \alpha_i^2 \lambda_i^2. \tag{26}$$

Teraz by sme chceli vyjadriť počet novoinfikovaných x_{n+1} v čase t_{n+1}. To môžeme ako počet nenulových zložiek vektora c, ktorý nám hovorí o počte nakazených susedov jednotlivých vrcholov (pričom v tomto jednoduchom prípade na infekciu stačí aspoň jeden nakazený sused) a ktorého zložky sú prirodzené čísla. Tento počet vieme ohraničiť pomocou vzťahov (24) a (26) nasledovne:

$$\begin{aligned} x_{n+1} = \#(c_i > 0) \leq \sum_{i=1}^{N} c_i \leq \sum_{i=1}^{N} c_i^2 = \\ = \sum_{i=1}^{N} \alpha_i^2 \lambda_i^2 \leq max|\lambda|^2 \sum_{i=1}^{N} \alpha_i^2 = max|\lambda|^2 x_n. \end{aligned} \tag{27}$$

Druhý spôsob ako ohraničiť tento počet je pomocou nerovnosti kvadratického a aritmetického priemeru:

$$\frac{c_1 + c_2 + \cdots + c_N}{N} \leq \sqrt{\frac{c_1^2 + c_2^2 + \cdots + c_N^2}{N}}. \tag{28}$$

Využitím vzťahov (24) a (26) dostávame ohraničenie:

$$\begin{aligned} x_{n+1} = \#(c_i > 0) \leq \sum_{i=1}^{N} c_i \leq \sqrt{N}\sqrt{c_1^2 + c_2^2 + \cdots + c_N^2} = \\ = \sqrt{N}\sqrt{\sum_{i=1}^{N} \alpha_i^2 \lambda_i^2} \leq \sqrt{N} max|\lambda| \sqrt{\sum_{i=1}^{N} \alpha_i^2} = \sqrt{N} max|\lambda| \sqrt{x_n}. \end{aligned} \tag{29}$$

Nás však vo všeobecnosti zaujímajú zmeny počtov novonakazených, ku ktorým ohraničenia budú nasledovné:

$$(27) \quad \Rightarrow \quad \frac{x_{n+1}}{x_n} \leq max|\lambda|^2 \tag{30}$$

$$(29) \quad \Rightarrow \quad \frac{x_{n+1}}{x_n} \leq max|\lambda| \sqrt{\frac{N}{x_n}}. \tag{31}$$

Ohraničenia (30) a (31) môžeme spojiť, čím dostávame finálnu nerovnosť:

$$\frac{x_{n+1}}{x_n} \leq min\left\{max|\lambda|^2, max|\lambda|\sqrt{\frac{N}{x_n}}\right\}. \tag{32}$$

Získali sme ohraničenie pre zmenu počtu novonakazených medzi časmi t_n a t_{n+1}. Pozrieme sa ešte na špeciálne prípady grafov, ich vlastnosti a čo z nich vyplýva pre zmenu počtu novonakazených.

Úplný graf

Teória spracovaná podľa [2].

Pre úplny graf s N vrcholmi platí, že spektrum jeho matice susedností je tvorené jednou vlastnou hodnotou $\lambda_1 = N - 1$ a ostatné vlastné hodnoty sú $\lambda_2 = \cdots = \lambda_N = -1$. Keby pracujeme v našom modeli s týmto špeciálnym grafom, máme prvého nakazeného v čase t_1, teda $x_1 = 1$. Keďže máme úplný graf, v druhom kroku by sa nakazili všetci ostatní, t. j. $x_2 = N - 1$. V takomto prípade by bola dĺžka jednej vlny pandémie 2 dni. Zmena počtu nakazených by bola:

$$\frac{x_{n+1}}{x_n} = \frac{N-1}{1} = N - 1 = max|\lambda|. \tag{33}$$

Pre tento špeciálny prípad by platila rovnosť zmeny počtu nakazených a maximálnej vlastnej hodnoty. V realite by to však znamenalo, že všetci ľudia sú navzájom v kontakte, čo nedáva žiaden zmysel.

Regulárny graf

Teória spracovaná podľa [2].

Graf sa nazýva regulárny so stupňom k (k-regulárny), ak všetky jeho vrcholy majú stupeň k. To znamená, že súčet v riadkoch aj v stĺpcoch jeho matice susedností je rovný k. Pre vlastné hodnoty k-regulárneho grafu platí, že ich absolútna hodnota je menšia, nanajvýš rovná ako stupeň k, t. j.

$$|\lambda_i| \leq k, \quad i = 1, \ldots, N \quad \Rightarrow \quad max|\lambda| \leq k. \tag{34}$$

Ak by sme predpokladali opäť prvého nakazeného v čase t_1, a teda $x_1 = 1$, ten by nakazil všetkých svojich k susedov, teda počet novonakazených v čase t_2 bude $x_2 = k$. Chceme vyjadriť zmenu počtu nakazených. Tá by v tomto prípade bola:

$$\frac{x_{n+1}}{x_n} = \frac{k}{1} = k \geq max|\lambda|. \tag{35}$$

Pre k-regulárny graf dostávame, že zmena počtu novonakazených je väčšia (nanajvýš rovná) ako v absolútnej hodnote maximálna vlastná hodnota. Takýto graf je však tiež veľmi vzdialený od reality a v skutočnosti pre zmeny počtov nakazených a maximálnu vlastnú hodnotu platí opačná nerovnosť. Predpoklad SIR modelov je, že všetky osoby sú rovnaké, a teda majú aj rovnaký stupeň vrcholov, čo predstavuje presne k-regulárny graf. Tento predpoklad modelov sa však ukazuje príliš zavádzajúci a skreslujúci realitu.

Súvislý a nesúvislý graf

Teória spracovaná podľa [2].

Nech λ_1 je najväčšia vlastná hodnota matice susedností grafu ($\lambda_1 = max\lambda$). Pre súvislý graf platí, ak je k-regulárny, tak $max\lambda = k$, inak $k_{min} < \bar{k} < max\lambda < k_{max}$, kde $k_{min}, \bar{k}, k_{max}$ sú najmenšia, priemerná a najväčšia vlastná hodnota. Pre nie nutne súvislý graf platí $\bar{k} \leq max\lambda \leq k_{max}$.

Ak by sme teraz v našom modeli uvažovali, že vrcholy novoinfikovaných sú náhodne vybrané, tak štatisticky môžeme predpokladať nasledujúci vzťah počtu novoinfikovaných v čase t_{n+1} a t_n:

$$\begin{aligned} x_{n+1} &\leq x_n\bar{k} \\ x_{n+1} &\leq x_n max\lambda \\ \frac{x_{n+1}}{x_n} &\leq max\lambda. \end{aligned} \tag{36}$$

Teda uvažujeme, že počet novoinfikovaných v čase t_{n+1} bude nanajvýš rovný počtu novoinfikovaných v čase t_n vynásobeného priemerným stupňom vrcholov. Nepredpokladáme, že sa približne rovnajú, keďže nerátame s opakovaniami, t. j. že novoinfikovaný mohol byť susedom viacerých infikovaných vrcholov. Režim opísaný predošlými nerovnosťami je režimom, ktorý môžeme pozorovať pri našom modeli, kde robíme viacero simulácií

a štatisticky nám platí nerovnosť (36). Toto ohraničenie je najsilnejšie zo všetkých a dokonca platí, že:

$$\frac{x_{n+1}}{x_n} \leq max\lambda \leq max|\lambda|. \tag{37}$$

Avšak neplatí všeobecne, ale iba za konkrétnych podmienok spomenutých vyššie, t. j. ak sú novoinfikovaní náhodne vybraní.

Máme dokázanú nerovnosť (32), ktorá platí pre všetky typy grafov. Všetky doteraz odvádzané nerovnosti a vzťahy však platia iba pre maticu susedností jednoduchého grafu, t. j. matica A_n obsahuje iba hodnoty 1 alebo 0. V našej práci prevažne uvažujeme vážený graf, ktorého jednotlivé prvky matice susedností nadobúdajú ľubovoľné hodnoty z intervalu $[0, 1]$. Čo sa zmení oproti predošlému postupu?

Vážený graf

Všetky vzťahy (21)-(26) platia aj v tomto prípade, avšak mení sa nám význam vektora c. V predošlom prípade hodnota c_i vyjadrovala počet infikovaných susedov vrchola i v čase t_n, pričom na infekciu stačila podmienka $c_i > 0$, aby sa vrchol i infikoval, ak ešte infikovaný nebol. Teraz nám hovorí o súčte pravdepodobností infikovania od jednotlivých vrcholov, t. j.:

$$c_i = \sum_{j \in J} p_j^i, \tag{38}$$

kde p_j^i je pravdepodobnosť infikovania sa vrcholu i od vrcholu j. Avšak ako sme ukázali v kapitole 4.2, tento prístup je nesprávny a pravdepodobnosť q_i nakazenia vrcholu i vypočítame ako:

$$q_i = 1 - \prod_{j \in J}^{N} (1 - p_j^i), \tag{39}$$

kde množina J je množinou indexov vrcholov, ktoré sú infikované (danej zložke vo vektore v prislúcha hodnota 1). Tento vzťah môžeme prepísať aj ako:

$$q_i = 1 - \left[1 - \sum_{j \in J} p_j^i + \sum_{j \in J} \sum_{\substack{k \in J \\ k \neq j}} p_j^i p_k^i - \dots \right] = \sum_{j \in J} p_j^i - \sum_{j \in J} \sum_{\substack{k \in J \\ k \neq j}} p_j^i p_k^i + \dots . \tag{40}$$

Definujme si novú konštantu k, pre ktorú platí:

$$k \gg c_i \left(1 - \frac{1}{x_n}\right). \tag{41}$$

Prečo sme si práve takto definovali konštantu si ukážeme pomocou nasledovných úprav:

$$\begin{aligned}
k &\gg c_i \left(1 - \frac{1}{x_n}\right) \\
\left(1 - \frac{1}{x_n}\right) c_i^2 &\ll k c_i \\
c_i^2 &\ll k c_i + \frac{c_i^2}{x_n} \\
c_i^2 &\ll k c_i + \sum_{j \in J} \left(p_j^i\right)^2 \qquad (42) \\
c_i^2 - \sum_{j \in J} \left(p_j^i\right)^2 &\ll k c_i \\
\sum_{j \in J} p_j^i \sum_{k \in J} p_k^i - \sum_{j \in J} \left(p_j^i\right)^2 &\ll k \sum_{j \in J} p_j^i \\
\sum_{j \in J} \sum_{\substack{k \in J \\ k \neq j}} p_j^i p_k^i &\ll k \sum_{j \in J} p_j^i. \qquad (43)
\end{aligned}$$

Úpravu na krok (42) sme realizovali na základe nerovnosti kvadratického a aritmetického rozdielu:

$$\begin{aligned}
\sqrt{\frac{\sum_{j \in J} \left(p_j^i\right)^2}{x_n}} \geq \frac{\sum_{j \in J} p_j^i}{x_n} &= \frac{c_i}{x_n} \\
\frac{\sum_{j \in J} \left(p_j^i\right)^2}{x_n} &\geq \frac{c_i^2}{x_n^2} \\
\sum_{j \in J} \left(p_j^i\right)^2 &\geq \frac{c_i^2}{x_n}.
\end{aligned}$$

Odvodená ostrá nerovnosť (43) nám hovorí o tom, že $k-$násobok prvého člena rozvoja (40) je oveľa väčší ako druhý člen tohto rozvoja. Ak zvolíme navyše $k \ll 1$, znamená to, že aj malý násobok prvého člena je stále oveľa väčší ako druhý člen tohto rozvoja.

Pozrieme sa, čo nám však táto podmienka vyjadruje:

$$\begin{aligned} k &\ll 1 \\ c_i\left(1-\frac{1}{x_n}\right) &\ll 1 \quad \text{(z definície (41))} \\ x_n\bar{p_i}\left(1-\frac{1}{x_n}\right) &\ll 1 \\ \bar{p_i}\left(x_n-1\right) &\ll 1, \end{aligned} \tag{44}$$

kde

$$\bar{p_i} = \frac{\sum_{j\in J} p_j^i}{|J|} = \frac{c_i}{x_n}. \tag{45}$$

Za podmienky (44) môžeme tvrdiť, že

$$\sum_{j\in J}\sum_{\substack{k\in J\\k\neq j}} p_j^i p_k^i \ll \sum_{j\in J} p_j^i \quad \Rightarrow \quad \frac{\sum_{j\in J}\sum_{\substack{k\in J\\k\neq j}} p_j^i p_k^i}{\sum_{j\in J} p_j^i} < \epsilon, \tag{46}$$

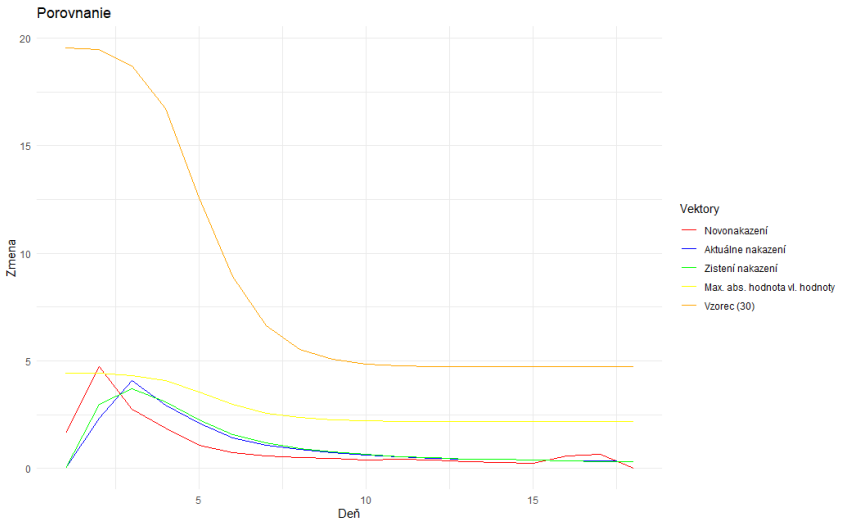
teda podiel druhého a prvého člena rozvoja je dostatočne malý.

Podobne by sme vedeli za danej podmienky ukázať, že hodnota členov v rozvoji (40) klesá, t. j. aj podiel tretieho a druhého člena bude dostatočne malý, tak isto podiel štvrtého a tretieho člena, atď. Hlavná myšlienka tohto tvrdenia spočíva v tom, že aj keď sa počet prvkov v sume daného člena zvýši, tak hodnoty týchto prvkov sa znižujú rádovo (predstavme si, že robíme súčiny pravdepodobností, ktoré sú rádovo $\bar{p_i} \approx 10^{-2}$). Čím v konečnom dôsledku za platnosti podmienky (44) môžeme aproximovať pravdepodobnosť nakazenia

$$\tilde{q_i} \approx c_i. \tag{47}$$

Za tejto aproximácie môžeme použiť ohraničenie (29), resp. ohraničenie pre zmenu (31). Vzťah (27) pre vážený graf použiť nemôžeme, keďže pre c_i definované podľa (38) vo všeobecnosti neplatí nerovnosť $\sum_{i=1}^{N} c_i \leq \sum_{i=1}^{N} c_i^2$, aj keď pre veľké N bude pre dostatočne hustú maticu táto nerovnosť platiť tiež.

Ukážeme si na príklade (obr.38), že odvodené ohraničenie (32) je až príliš benevolentné a pre reálne siete maximálnu zmenu počtu nakazených detekuje najlepšie maximálna absolútna hodnota vlastnej hodnoty.



Obr. 38: Porovnanie kriviek nakazených s maximálnou vlastnou hodnotou a ohraničením (32) (vlastné spracovanie v [6]).

5.3 Tri fázy priebehu epidémie

V počiatočnej fáze na začiatku epidémie, kedy máme ešte malý počet nakazených je veľmi malá pravdepodobnosť, že vektor nakazených sa zhoduje s maximálnym rastúcim vlastným vektorom matice susedností. Túto pravdepodobnosť by sme mohli matematicky vyjadriť ako:

$$P\left[|v - w_{max}| < \epsilon\right], \tag{48}$$

kde vektor v je vektor novonakazených, w_{max} je maximálny rastúci vlastný vektor prislúchajúci maximálnej vlastnej hodnote matice susedností a $\epsilon \ll 1$. Vyjadríme si vektor novonakazených v tejto fáze opäť ako lineárnu kombináciu vlastných vektorov w_i:

$$v = \sum_{i=1}^{N} \alpha_i w_i.$$

Za predpokladu (21) ortonormality vlastných vektorov matice susedností platí vzťah (24), t. j. $||v||^2 = \sum_{i=1}^{N} \alpha_i^2$. Pravdepodobnosť, že $|v - w_{max}| < \epsilon$ vieme ekvivalentne vyjadriť aj ako $\alpha_{max} \approx 1$.

Túto pravdepodobnosť by sme vedeli vyjadriť pomocou ϵ-okolia v N-rozmere, ako plochu povrchu nasledovne:

$$P\left[|v - w_{max}| < \epsilon\right] \approx \epsilon^{N-1}. \tag{49}$$

Kým je počet infikovaných stále nízky, matica susedností M sa zásadne nemení, čím sa veľmi nemení ani jej maximálna vlastná hodnota λ_{max} a k nej prislúchajúci vlastný vektor w_{max}. Postupne však pribúda počet nakazených a vektor novonakazených sa približuje k smeru maximálne rastúceho vlastného vektora. Nadobudnutím najväčšieho nárastu sa vektor novonakazených najviac priblíži už mierne modifikovanému maximálne rastúcemu vlastnému vektoru, čím aj doterajšia maximálna vlastná hodnota padne k nule a v matici susedností M vypadne radikálny počet prvkov.

Záver

Vírusové ochorenia predstavujú významnú hrozbu pre zdravie ľudí po celom svete. Vírusy môžu spôsobiť široké spektrum ochorení, od miernych infekcií až po vážne pandémie. Ich šírenie a závažnosť môžu byť ovplyvnené rôznymi faktormi, ako sú sociálne, ekonomické a environmentálne podmienky, ako aj účinnosť dostupných preventívnych opatrení a liečby. V súčasnosti sú vírusové ochorenia predmetom intenzívneho výskumu a monitorovania, najmä vzhľadom na ich schopnosť rýchlo sa šíriť a spôsobovať globálne zdravotné krízy. Cieľom našej práce je študovať priebeh simulovanej epidémie v dynamických reformátujúcich grafoch.

V prvej kapitole sme stručne charakterizovali jeden zo základných matematických modelov využívaných pri štúdiách šírenia infekčných chorôb. SIR model rozdeľuje populáciu do troch skupín a je matematicky reprezentovaný tromi diferenciálnymi rovnicami. Zaviedli sme definíciu reprodukčného čísla a z diferenciálnych rovníc odvodili jeho určitú formu.

V nasledujúcej kapitole sme si predstavili základné vlastnosti sietí, ich rozdelenie a metriky. Objasnili sme si aj pojem sociálna sieť, ako sieť ľudských kontaktov.

Tretia kapitola sa venuje základom teórie grafov. Vysvetlíme základné vlastnosti matice susedností, ktorá je základnou matematickou reprezentáciou grafu. Objasníme si rozdiely medzi rôznymi typmi grafov, ako jednoduchý graf, multigraf, či vážený graf. Jedným zo základných algoritmov používaných pri generovaní náhodných grafov je konfiguračný model. Vysvetlili sme jeho základné princípy fungovania.

Naša práca sa ďalej zamerala na vytvorenie modelu simulácie epidémie v prostredí R, ktorý nám umožnil simulovať rôzne scenáre. Najprv sme pomocou viacerých charakteristík vytvorili sieť reprezentujúcu ľudské kontakty. Prvým krokom bolo vytvorenie siete pomocou konfiguračného modelu a mocninového rozdelenia. Následne sme odstránili vlastné hrany a multihrany z grafu. Dopĺňaním hrán do trojuholníka sme zvýšili hustotu hrán v jednotlivých komponentoch, čím sme vytvorili akési skupinky ľudí. V poslednom kroku generovania grafu sme jednotlivým hranám grafu pridelili váhy na základe odvodenej distribúcie, čím sme rozlíšili silu kontaktov. Napokon ukazujeme princíp simulácie epidémie na takto vygenerovanom grafe.

V poslednej kapitole sme sa pozreli na konkrétne realizácie našich pozorovaní a analyzovali nami získané výsledky. Skúmali sme vzťah zmien počtu novonakazených, aktuálne nakazených a zistených nakazených, ktoré medzi sebou súvisia. Tiež sme analyzovali vzťah zmeny počtu nakazených a spektra matice susedností, jej maximálnej vlastnej hodnoty, resp. maximálnej absolútnej hodnoty vlastnej hodnoty. Ukázalo sa, že medzi touto zmenou a vlastnou hodnotou matice susedností platia rôzne vzťahy v závislosti od štruktúry grafu.

Výsledky našej práce nám poskytli hlboké pochopenie interakcie medzi dynamikou sociálnych sietí a šírením infekčných chorôb. Naša práca prispieva k lepšiemu pochopeniu problematiky šírenia infekčných chorôb prostredníctvom sociálnych sietí.

Zoznam použitej literatúry

[1] Boďová, K., Kollár, R.: Emerging Polynomial Growth Trends in COVID-19 Pandemic Data and Their Reconciliation with Compartment Based Models, arXiv: Populations and Evolution, 2020

[2] Brouwer, A. E., Haemers, W. H.: Spectra of graphs, Springer, Amsterdam, Netherlands, 2011, dostupné na internete (5.5.2024): https://www.win.tue.nl/ aeb/2WF05/spectra.pdf.

[3] Kwok, K. O., Cowling, B., Wei, V., Riley, S., Read, J. M.: Temporal variation of human encounters and the number of locations in which they occur: a longitudinal study of Hong Kong residents, The Royal Society, dostupné na internete (5.5.2024): http://dx.doi.org/10.1098/rsif.2017.0838.

[4] Newman, M.: Networks, Oxford university press, Spojené kráľovstvo, 2018.

[5] Psotová, S.: Fragmentácia náhodných sietí pri šírení infekčných chorôb a vakcinácii, Diplomová práca, FMFI UK, 2023

[6] R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2023, dostupné na internete (5.5.2024): https://www.R-project.org/.