

Středoškolská odborná činnost

Obor: 1. Matematika a statistika

Nový způsob simulace šíření epidemie ve společnosti

Autor: Jakub Dostál, Slovanské gymnázium Olomouc

Konzultant: Tomáš Füst, UP Olomouc

Olomouc 2016

Poděkování

Děkuji svému konzultantu Tomáši Fürstovi za obětavou pomoc při vedení mé práce. Nejvíce jsem mu však vděčný za nový pohled na vzdělání, matematiku a svět. Každá schůzka, rada, e-mail pro mne byla poučná, motivující a inspirující. Ukázal mi, že matematika a život se dá dělat i jiným způsobem.

Mým díkům nesmí uniknout ani jedna z osob, které mi pomáhaly s gramatickou korekturou a formální úpravou práce. Uznání však zaslouží všichni co snášejí mé nálady, způsobené mým osobním bojem s poznáním přírody jako takové.

V neposlední řadě děkuji projektu MŠMT 0017/7/NAD/2015 Rozvoj talentovaných žáků základních a středních škol prostřednictvím přírodovědných výzkumných aktivit a projektu Badatel.

Prohlašuji, že jsem tuto práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Prohlašuji, že tištěná verze a elektronická verze soutěžní práce SOČ jsou shodné.

V Olomouci dne

Title: New ways to simulate epidemic spreading in human society

Author: Jakub Dostál

School: Slavonic grammar school Olomouc

Supervisor: Tomáš Fürst, Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University Olomouc

Abstract:

Society needs to fight against dangerous diseases. However, disease may stand for much more abstract process, for instance a fashion wave.

The goal of the work is to show the need to separate the topology of the network of human relations from the behaviour of the disease itself. This will reveal us substantive dependence of the simulation results on the structure of society.

First basic ways of modelling epidemic spreading are shown, both deterministic and stochastic. Traditional methods of simulation on complex networks are recalled. Then a new algorithm is presented separating behaviour of the disease from the features of the network itself.

Keywords: epidemiology, complex networks, SI, SIS, SIR, SIRS

Název práce: Nový způsob simulace šíření epidemie ve společnosti

Autor: Jakub Dostál

Škola: Slovanské gymnázium Olomouc

Konzultant: Tomáš Fürst, Katedra matematické analýzy a aplikací matematiky, Přírodovědecká fakulta, Univerzita Palackého Olomouc

Abstrakt:

I v dnešní době společnost neustále pociťuje potřebu bojovat proti nebezpečným chorobám. Pod pojmem choroba si však můžeme představit i mnohem abstraktnější pojmy, jako je třeba módní trend.

Cílem práce je ukázat potřebu oddělení vlastností topologie vztahů lidí ve společnosti, kde se epidemie šíří, a samotného charakteru choroby. Tím zároveň ukážeme velkou závislost výpočtů právě na struktuře společnosti.

Nejprve uvádíme základní způsoby modelování, a to jak deterministické, tak stochastické. Ukazujeme tradiční způsoby simulací na komplexních sítích. Poté představujeme nový algoritmus oddělující vlastnosti sítě od vlastností samotné choroby.

Klíčová slova: epidemiologie, komplexní sítě, SI, SIS, SIR, SIRS

Obsah

1	Fundamentální deterministické modely	3
1.1	Model <i>SI</i>	3
1.2	Model <i>SIR</i>	5
1.3	Model <i>SIRS</i>	7
2	Stochastický přístup	9
2.1	Přechod k exponenciálnímu rozdělení	9
2.2	Stochastický přístup vs. determinismus	12
3	Modely na komplexních sítích	13
3.1	Klasifikace sítí	14
3.2	Síť světových letišť	15
3.3	Tradiční <i>SIS</i> algoritmus na komplexní síti	16
3.4	Nový algoritmus	17
4	Aplikace algoritmu	20
4.1	Simulace s konstantními vahami	21
4.2	Rozložení vah a simulace na ohodnocené síti	23
5	Závěr	25
	Reference	26
	Přílohy	27

Úvod

První vážné pokusy o modelování vývoje epidemií ve společnosti sahají do 18. století. Od té doby se mnohokrát změnil pohled na tuto problematiku, potřeba znalosti dynamiky chorob však stále přetrvává. V dnešním světě postiženém mnoha snadno přenosnými chorobami, jako je AIDS, různé druhy chřipek, nebo SARS, je stále nutné zkoumat do hloubky zákonitosti, které doprovází přenos těchto nemocí ve společnosti. Internet, cílená reklama, šíření politického přesvědčení, to vše jsou oblasti, na které se dají velmi snadno aplikovat znalosti nabyté právě studiem epidemií.

Směr a způsoby výzkumu v této oblasti se v průběhu dějin měnily. Na počátku práce vysvětlíme základní způsoby deterministického a stochastického modelování v nereálné dokonale mísené tekutině, kde každý jedinec ovlivňuje stejnou měrou jakéhokoliv jiného jedince. Poté se přesuneme na komplexní sítě, kam v dnešní době míří většina zájmu vědců. Pro komplexní sítě uvedeme nový způsob simulace, který odděluje vlastnosti sítě od vlastností samotné choroby. Pomocí něj dále ukážeme nutnost tohoto oddělení a vybrané vlastnosti *scale-free* sítí.

Všechny numerické výpočty a simulace byly prováděny v programovacím jazyku IPython [1] za použití jeho známých balíčků, například *scipy*, *numpy* nebo *matplotlib*.

1 Fundamentální deterministické modely

Studujeme-li šíření choroby ve společnosti, nemusíme se tolik zajímat o biologický pohled na nemoc. Není rozhodující, zda infikovaná osoba kašle, nebo má vyrážku. Zajímáme se pouze o stavy, v jakých se studovaný jedinec může nacházet. Nejdůležitější pro nás tedy je, zda je osoba imunní, infekční, nebo třeba mrtvá. Tato abstrakce nám dává jednotný pohled na mnohem více, než jen lidská onemocnění. Můžeme studovat počítačové viry, rozvoj politických myšlenek nebo zpráv ve společnosti pomocí stejných nebo velmi podobných metod. Slovo „choroba“ od této chvíle tedy budeme chápat v mnohem širším smyslu. V celé práci budeme pro jednoduchost užívat převážně biologické názvosloví i při popisu nebiologických systémů.

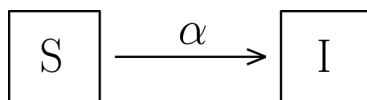
První, kdo se pokoušel matematicky modelovat epidemie, byl fyzik Daniel Bernoulli [2]. V roce 1766 přišel se systémem popsáním diferenciálními rovnicemi. Takovéto rovnice nám dovolují nahlížet na šíření nemoci jako na chemickou reakci v dokonale míšené kapalině.

Před tím, než se začneme bavit o konkrétních způsobech modelování epidemií, bychom je měli roztrdit do několika kategorií podle jejich průběhu. Nejprve definujeme stavy, ve kterých se osoba může nacházet. Na začátku je většina jedinců *zdravá* (*susceptible*, S). V tomto stavu nemůžeme nijak ovlivnit žádnou jinou osobu, my sami však můžeme být nakaženi. Když nás někdo nakazí, stáváme se *infekčními* (*infected*, I) a můžeme nakazit ostatní jedince. V některých modelech se po nějakém čase dostaneme do fáze označované písmenem R . Interpretace tohoto stavu se liší napříč jednotlivými modely. Někdy tím myslíme *vyléčený* (*recovered*), někdy *imunní* (*resistant*), jindy *mrtvý* (*dead*). V každém případě však v tomto stavu nemůžeme nikoho ovlivnit a zároveň nikdo nemá jak ovlivnit nás.

Když nyní známe základní stavy, se kterými budeme pracovat, nic nám nebrání udělat krok vpřed a definovat několik modelů. Existují celkem čtyři základní modely (SI , SIS , SIR , $SIRS$), které pracují pouze se čtyřmi stavy definovanými výše. Tyto procesy jsou popsány diferenciálními rovnicemi, ale je snadné napsat algoritmy, které děje modelují přirozeněji než rovnice (společnost není dokonale míšená kapalina). Jinými slovy, naše níže popsané modely jsou pouze popisy způsobu šíření choroby mezi lidmi a může na ně být nahlíženo z více úhlů pohledu. Plně vysvětlíme chování jen pro některé modely, s jejichž znalostí si již snadno představíme modely zbývající.

1.1 Model SI

Tento model říká, že se nemoc může přenášet pouze z infikovaných jedinců na zdravé jedince. Žádnou jinou reakci nepovoluje. Obrázek 1 popisuje graficky tento proces. Rychlost tohoto přenosu závisí na parametru α konkrétní choroby. Inter-



Obrázek 1: Diagram SI šíření choroby.

pretace parametru závisí na úhlu našeho pohledu (algoritmus, dif. rovnice).

Z biologického hlediska můžeme říct, že jakmile je člověk infikován, nikdy se nevyhlídí a v každém okamžiku může nakazit kohokoliv jiného. Tento model odpovídá například šíření informace na základní škole o tom, že Ježíšek neexistuje.

Tento model můžeme snadno reprezentovat systémem dvou diferenciálních rovnic. Nejprve definujeme počáteční podmínky. Necht' N je počet studovaných lidí, nikdo nepřichází ani neodchází. V čase $t = 0$ je zde I_0 infikovaných a S_0 zdravých. Jelikož náš model pracuje pouze s těmito dvěma stavy, je zřejmé, že $S_0 + I_0 = N$. Změní-li se počet zdravých jedinců, musí se stejně velká změna stát i u nakažených jedinců. Definujme funkce času $S(t), I(t)$, kde $S(t)$ je počet *zdravých* jedinců a $I(t)$ počet infikovaných jedinců v čase t .¹ Nyní můžeme psát:

$$\frac{dS}{dt} + \frac{dI}{dt} = 0. \quad (1.1)$$

Podobně jako při chemických reakcích je rychlost přenosu nemoci přímo úměrná počtu nakažených i počtu zdravých osob (jejich obdobou budou v chemických reakcích koncentrace látek). Čím bude více zdravých a infekčních, tím rychleji se nemoc bude přenášet. Nyní již snadno napíšeme rovnice, které toto popisují:

$$\begin{aligned} \frac{dS}{dt} &= -aSI \\ \frac{dI}{dt} &= aSI, \end{aligned} \quad (1.2)$$

kde a je parametr určující rychlost šíření choroby.² Vzhledem k tomu, že $S(t) + I(t) = N \Rightarrow S(t) = N - I(t)$, můžeme systém 1.2 přepsat do snadno řešitelné podoby a najít řešení pouze pro $I(t)$. Jelikož celkový počet studovaných jedinců N se nemění, počet zdravých osob v čase t bude $S(t) = N - I(t)$.

$$\frac{dI}{dt} = aI(N - I) \quad (1.3)$$

Nyní vydělíme celou rovnici 1.3 celkovým počtem lidí N . Definujeme novou proměnnou $i(t) := \frac{I(t)}{N}$ a parametr $\alpha := aN$.

$$\frac{di}{dt} = \alpha i(1 - i) \quad (1.4)$$

Nyní separujeme proměnné a dostáváme:

$$\frac{1}{i(1 - i)} di = \alpha dt. \quad (1.5)$$

Zde rozložíme na parciální zlomky a integrujeme. Díky naší epidemiologické motivaci víme, že $\forall t : i(t) \in (0, 1)$ a proto můžeme snadno psát:

$$\ln(i) - \ln(1 - i) = \alpha t + c. \quad (1.6)$$

¹Pro jednoduchost budeme, pokud neuvedeme jinak, bez dalších připomínek užívat písmena S, I jako funkce času definované výše.

²Toto je analyticky snadno řešitelná soustava, rovnice jiných modelů však tak snadné nejsou. Většinou analytické řešení vůbec nemají.

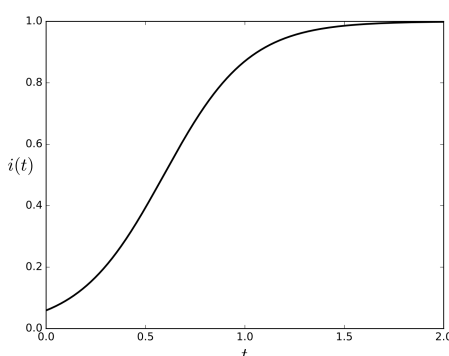
Po několika algebraických úpravách nacházíme pěkný vztah pro $i(t)$.

$$i(t) = \frac{e^{\alpha t+c}}{1 + e^{\alpha t+c}} \quad (1.7)$$

Nyní po porovnání 1.7 s počátečními podmínkami snadno dostaneme hodnotu integrační konstanty c , a tedy i přesný počet infikovaných osob v jakémkoliv čase t :

$$i(t) = \frac{i_0 e^{\alpha t}}{i_0 (e^{\alpha t} - 1) + 1}, \quad (1.8)$$

kde $i_0 = \frac{I_0}{N}$.



Obrázek 2: Numerické řešení rovnice 1.3 pro $\alpha \doteq 4.7$ a $i_0 \doteq 0.06$.

Model SI není příliš komplikovaný. Jak můžeme vidět z grafu 2, po nějaké době je většina lidí infikovaná. Můžeme snadno ukázat, že pro $t \rightarrow \infty$ model předpovídá $i(t) \rightarrow 1$. To znamená, že po dostatečně dlouhé době je každý ve studované skupině infikovaný.

1.2 Model SIR

Model SIR je trochu více komplikovaný než model SI , ale hodí se pro jiné choroby, například neštovice. Proces infikování a možnosti *zdravých* jedinců jsou stejné jako v modelu SI . Jediný rozdíl je v tom, že po nějaké době se *infikovaní* jedinci stávají *mrtvými/imunními* – jednoduše už nijak neovlivňují studovanou společnost. Tento proces závisí na parametru β . Graficky je znázorněn diagramem 3.



Obrázek 3: Diagram SIR šíření nemoci.

Není důležité jak pojmenujeme stav R , jediná rozhodující skutečnost je, že v tomto stavu jedinec nemůže nikoho ovlivnit a nikdo nemůže nijak ovlivnit jeho.

Můžeme snadno odvodit rovnice popisující tento proces. Použijeme stejnou notaci jako u SI modelu, kterou intuitivně rozšíříme o několik nových pojmů.

Je zřejmé, že pro SIR bude platit podobný vztah jako pro SI platí 1.1, tedy že v průběhu pozorování je celkový počet osob konstantní.

$$\frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = 0. \quad (1.9)$$

Na rozdíl od SI máme tři možné stavy, tedy potřebujeme tři rovnice. Stejně tak je nutné snižovat počet osob ve stavu I a zvyšovat počet osob ve stavu R rychlostí, která závisí na počtu *infikovaných* osob a parametru β . Náš systém tedy vypadá takto:

$$\begin{aligned} \frac{dS}{dt} &= -aSI \\ \frac{dI}{dt} &= aSI - \beta I \\ \frac{dR}{dt} &= \beta I. \end{aligned} \quad (1.10)$$

Problémem je, že tato soustava diferenciálních rovnic není analyticky řešitelná. I přesto však můžeme odvodit některé zajímavé vlastnosti systému. Nejprve přepíšeme soustavu 1.10 do užitečnější podoby. Vydělíme ji celkovým počtem studovaných lidí N a definujeme nové proměnné $s(t) = \frac{S(t)}{N}$, $i(t) = \frac{I(t)}{N}$, $r(t) = \frac{R(t)}{N}$ a parametr $\alpha = aN$. Parametr β zůstává stejný.

$$\begin{aligned} \frac{ds}{dt} &= -\alpha si \\ \frac{di}{dt} &= \alpha si - \beta i \\ \frac{dr}{dt} &= \beta i. \end{aligned} \quad (1.11)$$

Je zřejmé, že pro $i_0 = 0$ je $s(t) = s_0$, $i(t) = i_0 = 0$, $r(t) = r_0$. Proto uvažujme $i_0 > 0$ a $r_0 = 0$ ³. Díky naší biologické motivaci víme, že $\alpha > 0$, $\beta > 0$. Nyní, je zřejmé, že $\frac{ds}{dt} < 0$, $\frac{dr}{dt} > 0$, proto je počet *zdravých* jedinců vždy klesající a počet *imunních* jedinců vždy rostoucí. Zákonitosti pro $\frac{di}{dt} = \alpha si - \beta i$ jsou více zajímavé. Můžeme vidět, že pro $s > \frac{\beta}{\alpha}$ je počet *infikovaných* rostoucí, pro $s = \frac{\beta}{\alpha}$ konstantní a pro $s < \frac{\beta}{\alpha}$ klesající. To znamená, že je-li počáteční počet *zdravých* jedinců $s_0 < \frac{\beta}{\alpha}$, potom $i(t)$ je stále klesající. To vše je pěkně vidět z fázového portréту proměnných s a i . Můžeme snadno odvodit jeho exaktní tvar z rovnice 1.11. Ekvivalentními úpravami ji transformujeme do podoby:

$$\begin{aligned} dt &= -\frac{ds}{\alpha s i} \\ dt &= \frac{di}{\alpha s i - \beta i}, \end{aligned} \quad (1.12)$$

dále:

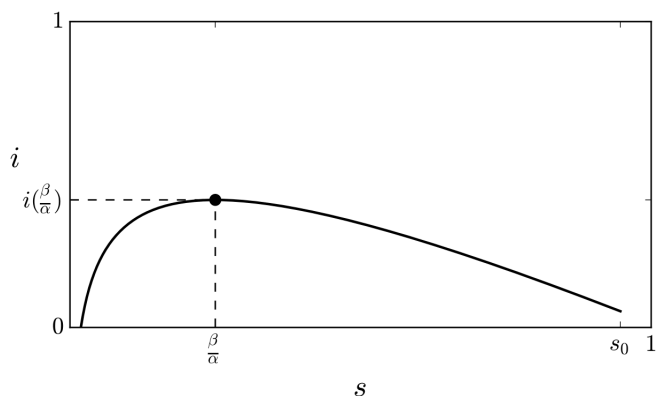
$$-\frac{ds}{\alpha s i} = \frac{di}{\alpha s i - \beta i}. \quad (1.13)$$

³Imunní jedinci neovlivňují nijak systém, můžeme tedy bez újmy na obecnosti položit $r_0 = 0$.

Po separování proměnných, integraci a nalezení integrační konstanty z počátečních podmínek dostáváme implicitní předpis pro $i(s)$, tedy náš fázový portrét.

$$i(s) = \frac{\beta}{\alpha} \ln(s) - s - \frac{\beta}{\alpha} \ln(s_0) + 1 \quad (1.14)$$

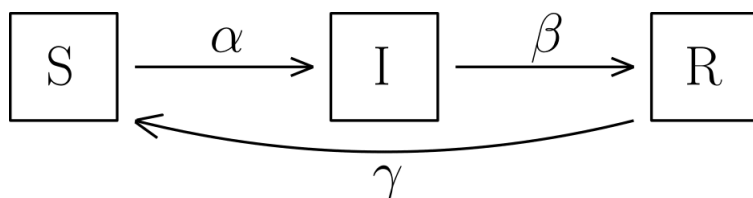
Z fázového portréту 4, vidíme, že existuje lokální maximum. Po porovnání $\frac{di}{ds}$ s nulou snadno dokážeme, že se nachází v bodě $s = \frac{\beta}{\alpha}$. Z rovnice 1.14 můžeme spočítat přesnou hodnotu $\max(i) = i(\frac{\beta}{\alpha})$, ta však nemá příliš hezký tvar.



Obrázek 4: *SIR* fázový portrét pro $\alpha = 0.8$, $\beta = 0.2$ a $i_0 = 0.053$.

1.3 Model *SIRS*

Model *SIRS* je o něco komplikovanější než *SIR*, ale více se hodí pro jiné choroby, například pro chřipku. V diagramu 5 můžeme pozorovat chování tohoto modelu. Říká nám, že můžeme být v jednom ze tří stavů: *zdravý*, *infekční* nebo *imunní*.



Obrázek 5: Diagram *SIRS* šíření nemoci.

Proces infikování je stejný jako u *SIR* modelu, ale po nějaké době, kdy jsme *imunní*, se znovu stáváme *zdravými* a nakazitelnými. Tato transformace závisí na parametru γ .

Stejně snadno jako pro *SIR* a *SI* modely můžeme odvodit systém diferenciálních rovnic i pro *SIRS*. Vyjděme z rovnice 1.10 pro *SIR* a přidejme změnu z *R* do *S*, která závisí na počtu *imunních* jedinců a parametru γ . Přepišme však rovnou celou soustavu do podoby po vydělení celkovým počtem sledovaných jedinců:

$$\begin{aligned} \frac{ds}{dt} &= \gamma r - \alpha si \\ \frac{di}{dt} &= \alpha si - \beta i \\ \frac{dr}{dt} &= \beta i - \gamma r. \end{aligned} \quad (1.15)$$

Stejně jako v případě *SIR* toto není analyticky řešitelný systém. Můžeme o něm však říct některé užitečné informace.

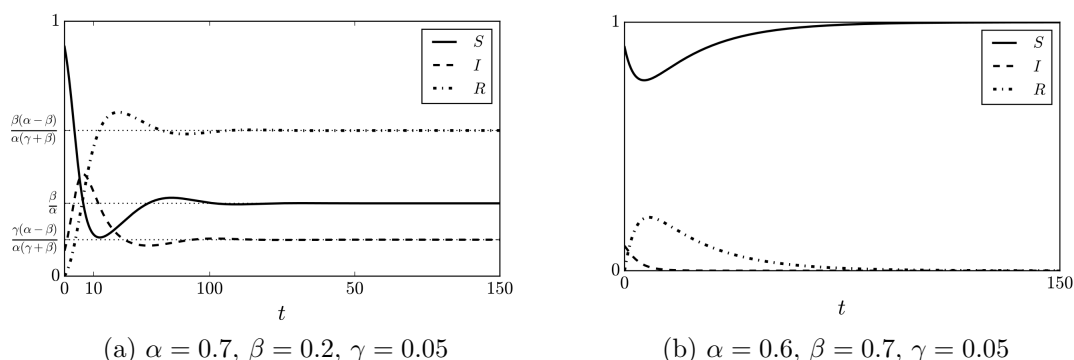
Tento systém má jeden triviální stabilní stav $i = 0$. Je zde však zajímavější netriviální stabilní stav. Položíme-li $\frac{ds}{dt} = \frac{di}{dt} = \frac{dr}{dt} = 0$ a uvážíme, že $s(t) + i(t) + r(t) = 1$, obdržíme následující soustavu nelineárních algebraických rovnic:

$$\begin{aligned}\gamma r - \alpha s i &= 0 \\ \alpha s i - \beta i &= 0 \\ \beta i - \gamma r &= 0.\end{aligned}\tag{1.16}$$

Po jejím vyřešení dostáváme netriviální stacionární řešení:

$$\begin{aligned}s &= \frac{\beta}{\alpha} \\ i &= \frac{\gamma(\alpha - \beta)}{\alpha(\gamma + \beta)} \\ r &= \frac{\beta(\alpha - \beta)}{\alpha(\gamma + \beta)},\end{aligned}\tag{1.17}$$

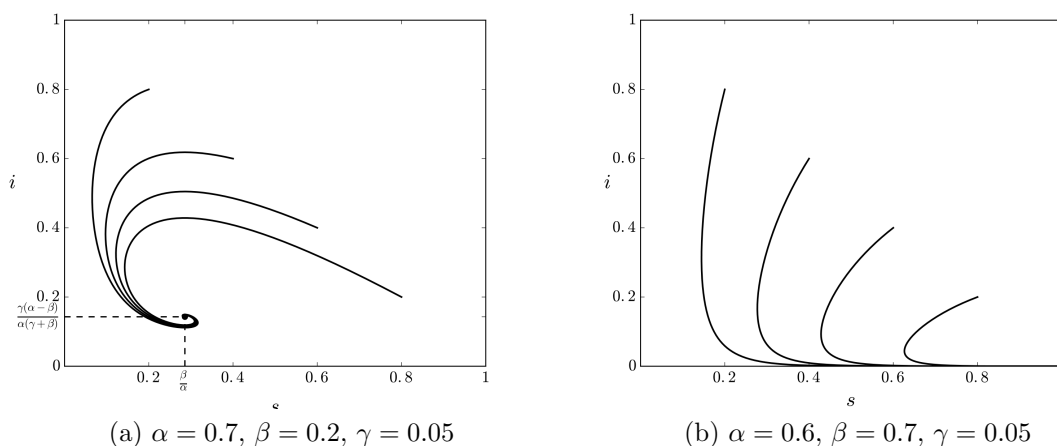
které názorně ukazuje graf 6(a). Uvědomme si, že soustavu jsme řešili jen pro



Obrázek 6: Numerické řešení *SIRS* systému 1.16 pro $i_0 = 0.1, r_0 = 0$.

takové α, β, γ , pro které se hodnoty s, i, r nachází v intervalu $(0, 1)$. Uvažujeme-li totiž $\beta > \alpha$, podle 1.17 by jistě $s > 1$ a $i < 0$, což očividně není možné. Je však zřejmé, že pro $\beta > \alpha$ nemoc vymírá, tedy $s \rightarrow 1$ a zároveň $i \rightarrow 0$. To můžeme vidět v grafu 6(b).

Zajímavé informace lze zjistit také z fázového diagramu proměnných i a s . Bohužel jej není možné odvodit analyticky, lze jej však snadno numericky dopočítat. V grafech 7 vidíme fázové portréty se stejnými parametry jako jsou v grafu 6, avšak pro různé počáteční podmínky. Povšimněme si, že výsledný stav systému skutečně nezávisí na hodnotách s_0, i_0 . Na obou obrázcích jsou diagramy postupně pro $i_0 = 0.2, i_0 = 0.4, i_0 = 0.6, i_0 = 0.8$. V případě $\alpha > \beta$, tedy 7(a), se hodnoty s, i dostávají do stacionárního stavu se souřadnicemi $s = \frac{\beta}{\alpha}, i = \frac{\gamma(\alpha - \beta)}{\alpha(\gamma + \beta)}$. V případě, kdy je $\beta > \alpha$, tedy 7(b), nemoc vymírá a směřuje do stavu $s = 1, i = 0$.



Obrázek 7: Fázový portrét *SIRS* systému 1.16 pro $r_0 = 0$ a i_0 postupně $i_0 = 0.2, i_0 = 0.4, i_0 = 0.6, i_0 = 0.8$.

2 Stochastický přístup

Je jistě zřejmé, že šíření nemoci ve společnosti nefunguje stejně deterministicky jako systém popsany obyčejnými diferenciálními rovnicemi představený v předchozí kapitole. Lidé se nepotkávají po přesně daných intervalech času, ani není jisté, že při každém setkání si chorobu předají. Stejně tak se může stát, že potkáme naprosto náhodně jedince, který nás nakazí, aniž bychom s ním byli v nějakém pravidelném kontaktu. To byl jeden z hlavních důvodů vzniku nových pravděpodobnostních implementací základních modelů.

2.1 Přechod k exponenciálnímu rozdělení

Pro jednoduchost a pochopení základních principů nejprve vytvořme simulaci v dokonale míseném prostředí, tedy bez jakýchkoliv ohledů na topologii vztahů společnosti.

Pokusme se tedy nějakým způsobem zakomponovat náhodu do našeho modelu *SIS*. V textu níže budeme dále definovat pravděpodobnosti několika událostí, které označíme malými písmeny p s různými indexy. Pod zápisem s velkým písmenem $P[A]$ však chápeme pravděpodobnost obecného jevu A . Zavedme proměnné $S(t), I(t), R(t)$ stejným způsobem jako dříve. Uvažme malý element času dt . Víme, že $I(t)$ je počet nakažených v čase t . Potom stav systému v čase $t + dt$ může být následující⁴:

1. pokud dojde k jedné nákaze: $I(t + dt) = I(t) + 1$, tuto události budeme značit \mathcal{I}^+ ;
2. pokud dojde k jednomu uzdravení⁵: $I(t + dt) = I(t) - 1$, tuto události budeme značit \mathcal{I}^- .

⁴Stejným způsobem funguje náhodná procházka na 1D ose.

⁵Uzdravením v *SIS* modelu je myšlen přechod ze stavu I zpět do stavu S .

Žádný jiný děj v *SIS* modelu nemůže pro dostatečně malé dt proběhnout. Všimněme si, že zatím neřešíme, kdo se uzdravil, nebo nakazil. Uvažujeme, že se nakazila $1/N$ celkové skupiny sledovaných jedinců, kde N je počet osob.

Hledejme nyní pravděpodobnost, že za čas dt dojde k nákaze, tedy k ději \mathcal{I}^+ , označme ji p^+ . Aby došlo k nákaze, musí se setkat nakažený jedinec se zdravým jedincem. Pravděpodobnost, že tato situace nastane, je jistě úměrná „koncentracím“ jednotlivých stavů ve sledovaném souboru. Pod tímto chemickým názvoslovím se neskrývá nic jiného než podíl, kterým je konkrétní stav zastoupen ve zkoumaném systému, tedy $\frac{S}{N}$ a $\frac{I}{N}$. Nezapomínejme, že S a I jsou funkce času, tento podíl tedy nebude v čase konstantní. Zavedme opět obdobné značení pro koncentrace jako v první kapitole: $s(t) := \frac{S(t)}{N}$, $i(t) := \frac{I(t)}{N}$. Pravděpodobnost, že se dva jedinci s těmito stavy potkají, je $\frac{S}{N} \cdot \frac{I}{N} = si$. Je zřejmé, že p^+ bude dále úměrné elementu času dt a přenosové rychlosti α . Ta v tomto případě představuje pravděpodobnost, že se infekce při setkání nakažitelného a infekčního přenesla za čas dt . Má tedy jednotku $[\alpha] = 1/s$. Nyní již snadno dostáváme $p^+(t)$ tedy pravděpodobnost, že mezi časem t a $t + dt$ se jedna osoba nakazí⁶

$$p^+(t) = P[\mathcal{I}^+] = dt \alpha s(t) i(t). \quad (2.1)$$

Obdobně odvodíme pravděpodobnost, že se mezi časem t a $t + dt$ jedinec uzdraví $p^-(t)$, tedy proběhne děj \mathcal{I}^- . Tento děj závisí na parametru β , s obdobnou interpretací a stejnou jednotkou jako α a koncentraci nakažených $\frac{I}{N} = i$.

$$p^-(t) = P[\mathcal{I}^-] = dt \beta i(t) \quad (2.2)$$

Jelikož v modelu *SIS* uvažujeme jen tyto dva děje, můžeme snadno dopočítat pravděpodobnost, se kterou se mezi časem t a $t + dt$ nic nestane, tedy nikdo nebude nakažený a nikdo se neuzdraví:

$$p^{nic}(t) = 1 - p^+(t) - p^-(t). \quad (2.3)$$

Pokusme se nyní najít pravděpodobnost, že po dobu T se nic dít nebude a ihned potom v čase $t + T$ se něco stane, tedy někdo se nakazí nebo uzdraví. Tuto pravděpodobnost označme $p(T)$. Uvědomme si však, že závisí i na počátečním čase t na kterém implicitně závisí koncentrace. Pro jednoduchost však tento parametr psát nebudeme. Jelikož jsou děje \mathcal{I}^+ a \mathcal{I}^- disjunktní, pravděpodobnost, že se stane alespoň jeden z nich je $P[\mathcal{I}^+ \vee \mathcal{I}^-] = (p^+ + p^-)$. Rozdělme si časový úsek T na M elementů času Δt , tedy $T = M\Delta t$. Nyní tedy M časových kroků zůstává systém ve stejném stavu a poté se stane jedna z možných událostí. To znamená, že se M -krát stane děj s pravděpodobností p^{nic} . Jelikož jde o M nezávislých dějů, které však musí všechny nastat v konjunkci, pravděpodobnost, že se tak stane je $(p^{nic})^M$. Dále tedy již snadno:

$$\begin{aligned} p(T) &= (p^{nic})^M (p^+ + p^-) = \\ &= (1 - p^+ - p^-)^M (p^+ + p^-). \end{aligned} \quad (2.4)$$

⁶Musíme si uvědomit, že tato pravděpodobnost závisí implicitně na čase, neboť je závislá na koncentracích, které jsou proměnnými času, proto píšeme $p^+(t)$. U koncentrací, tedy hodnot $s(t), i(t)$, nebudeme pro jednoduchost vždy argument s časem psát.

Pro přehlednost označme $\omega(i) = \alpha si + \beta i = \alpha i(1 - i) + \beta i$. Je funkcí koncentrace nakažených, tedy funkcí času⁷. Nyní substituuje rovnici 2.4 s rovnicemi 2.1, 2.2 a dostáváme:

$$\begin{aligned} p(T) &= (1 - \Delta t \alpha si - \Delta t \beta i)^M (dt \alpha si + dt \beta i) = \\ &= (1 - \omega(i) \Delta t)^M \omega(i) dt \end{aligned} \quad (2.5)$$

Přepíšeme-li nyní element Δt v závorce podle našeho dělení času T , dostáváme:

$$p(T) = \left(1 - \frac{\omega(i)T}{M}\right)^M \omega(i) dt. \quad (2.6)$$

Z tohoto tvaru již snadno [3] limitně pro $M \rightarrow +\infty$:

$$\lim_{M \rightarrow +\infty} p(T) = dt \omega(i) e^{-\omega(i)T}. \quad (2.7)$$

Naše pravděpodobnost $p(T)$ toho, že po čas T se nic dít nebude a poté se ihned něco stane, má tedy zřejmě exponenciální rozdělení s parametrem ω závislým na koncentraci infekčních jedinců.

Tato znalost nám dovoluje sestavit jednoduchý algoritmus simulující šíření nemoci v dokonale promíchané společnosti. Tím, že je postaven na náhodných událostech, je opět o něco blíže skutečnosti, než jsou deterministické modely uváděné v předchozí kapitole.

Pro pravděpodobnosti $p^+(t), p^-(t)$ platí, že $p^+(t) + p^-(t) < 1$. Nyní však víme, že mezi časem T a časem $T + dt$ jeden z dějů nastane určitě. Zavedme tedy nové pravděpodobnosti $q^+(t), q^-(t)$, pro které musí být $q^+(t) + q^-(t) = 1$. Dále chceme zachovat vzájemný poměr původních pravděpodobností, platí tedy i $\frac{p^+(t)}{p^-(t)} = \frac{q^+(t)}{q^-(t)}$. Nyní můžeme snadno tyto dvě rovnice řešit pro neznámé $q^+(t), q^-(t)$. Výsledné hodnoty budou zřejmě naše hledané pravděpodobnosti.

Data: Počáteční podmínky, tj. N, I_0 , parametry α, β , dále počáteční čas $t_0 = 0$ a čas t_{max} , kdy chceme simulaci ukončit.

```

1  $t := t_0$ ;
2 while  $t < t_{max}$  do
3   | vybereme čas  $T$  z exponenciálního rozdělení s parametrem  $\omega(i)$ ;
4   | vybereme náhodné  $x \in (0, 1)$ ;
5   | if  $x < q^+(t)$  then
6     |    $s := s - 1/N$ ;
7     |    $i := i + 1/N$ ;
8   | else
9     |    $s := s + 1/N$ ;
10    |    $i := i - 1/N$ ;
11  | end
12  |  $t := t + T$ 
13 end
```

Algoritmus 1: Algoritmus *SIS* simulace.

Tento algoritmus simuluje šíření choroby typu *SIS* ve společnosti. Příkazy na řádcích 4, 5, 8 zajišťují pouze výběr mezi tím, zda bude jeden jedinec infikován, nebo vyléčen.

⁷Pro přehlednost však budeme užívat i tvar bez argumentu $\omega(i)$.

Celkem snadno bychom mohli užívat algoritmus, ve kterém bychom každý časový element dt vybrali jednu ze tří možných variant p^+, p^-, p^{nic} . Problém je však v tom, že $p^{nic}(t) \gg p^+(t)$, a samozřejmě také $p^{nic}(t) \gg p^-(t)$. V drtivé většině případů bychom tedy vybrali p^{nic} . Proto nám přechod na exponenciální rozdělení pro velmi malé dt výrazně šetří strojový čas.

Obdobným způsobem by se daly vytvořit algoritmy i pro ostatní modely. Naším cílem však bylo ukázat odvození výběru tzv. waiting time z exponenciálního rozdělení.

2.2 Stochastický přístup vs. determinismus

To, že stochastický přístup je o něco reálnější než ten deterministický, je snad jasné již z úvodu této kapitoly. Nyní však ukážeme, že provádíme-li stochastické simulace v dokonale mísené společnosti, jsme schopni jej s dostatečnou přesností přirovnat k deterministickému modelu. Tím snad přiblížíme nutnost modelování na komplexních sítích, kterou osvětlíme v kapitole následující.

Pro jednoduchost budeme dále pracovat s modelem SI představeném v sekci 1.1 a algoritmem 2 představeným níže.

Data: Počáteční podmínky, tj. N, I_0 , parametr α , dále počáteční čas $t_0 = 0$.

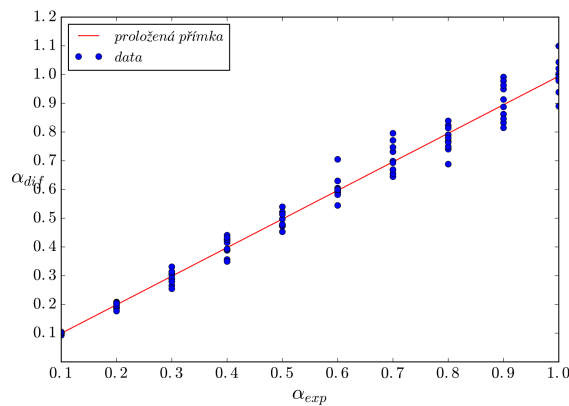
```

1  $t := t_0$ ;
2 while  $i < 1$  do
3   | vybereme čas  $T$  z exponenciálního rozdělení s parametrem  $\omega(i)$ ;
4   |  $i := i + 1/N$ ;
5   |  $t := t + T$ 
6 end

```

Algoritmus 2: Algoritmus SI simulace.

Tímto algoritmem jsme spočítali celkem 100 simulací, vždy deset simulací pro deset hodnot argumentu $\alpha_{exp} = 0.1, 0.2, \dots, 1.0$. Počáteční podmínky pro všechny simulace byly $N = 1000, I_0 = 10$. Dále jsme simulovaná data fitovali pomocí metody nejmenších čtverců rovnicí 1.8. Tím jsme našli jednotlivé hodnoty parametrů α_{dif} deterministického systému, které nejvíce vyhovují parametrům daným ve stochastickém modelu.

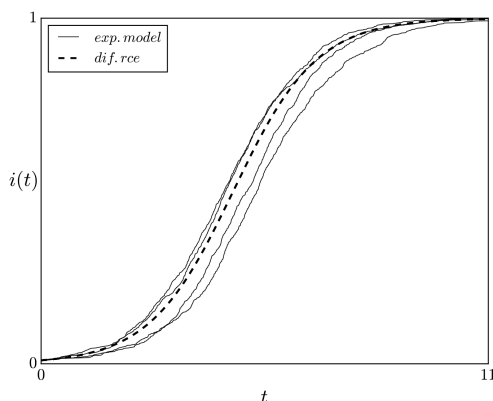


Obrázek 8: Závislost parametrů deterministického a stochastického modelu.

Na grafu 8 jsou na vodorovné ose jednotlivé hodnoty parametrů α_{exp} , které byly použity při výpočtu pomocí algoritmu 2. Tečky jsou nejvhodnější parametry α_{dif} diferenciálního systému pro jednotlivé běhy algoritmu nalezené metodou nejmenších čtverců.

Můžeme si povšimnout, že pro zvyšující se α_{exp} se zvyšuje i rozptyl nalezených hodnot α_{dif} . Relativní přesnost je konstantní, když však α_{exp} roste, zvětšuje se absolutní přesnost, tedy rozptyl.

Jde vidět, že nalezené hodnoty si dobře odpovídají, nemají velký rozptyl. Jde tedy říci, že diferenciální systém a náš stochastický algoritmus budou v tomto případě dávat podobné výsledky. O tom se můžeme přesvědčit i z grafu 9 závislosti zlomku infekčních jedinců na čase.



Obrázek 9: Graf závislosti zlomku infekčních jedinců na čase ve stochastickém a deterministickém pojetí modelu SI .

I přesto jsou stochastické modely v dokonale míseném prostředí o něco reálnější než deterministické modely. Nejvíce se to ukazuje při modelování chorob typu SIS , nebo $SIRS$, které nejsou v čase monotónní. Ukazuje se, že v takových případech může choroba úplně vymřít, což nám deterministické modely nedovolují. Úkolem této kapitoly je však ukázat, že ani modely zahrnující náhodu nejsou dokonalé. Pro reálnější výsledky bychom se měli přesunout z dokonale míseného prostředí na komplexní sítě s reálnější topologií vztahů.

3 Modely na komplexních sítích

Komplexní systém si můžeme představit jako množinu mnoha bodů, kde některé body jsou navzájem propojeny. Když se podíváme na mapu, každé město označíme jako bod (tečku) a každou silnici spojující města zakreslíme rovnou přímkou, dostáváme komplexní síť. Stejně tak mohou fungovat křižovatky a silniční síť, propojení počítačů v LAN síti, nebo například letiště a přímé spoje mezi nimi. Komplexní sítě mohou představovat ale i o něco abstraktnější pojmy. Můžeme zkoumat vědecké články a citace v nich uváděné, World Wide Web a odkazy na jednotlivých stránkách. Zajímavou sítí je také systém přátelství mezi lidmi, kde jedinec je vrchol a pravidelný kontakt hrana sítě.

Jak je zřejmé z předchozího odstavce, při studiu komplexních sítí je hojně užíváno matematických poznatků a názvosloví teorie grafů. I my budeme v další

části textu často zaměňovat například slovo *jedinec* se slovem *vrchol*. Stejně tak budeme pod *hranou* chápat vzájemný kontakt dvou *vrcholů/jedinců*.

Přesněji řečeno, komplexní síť je tedy množina vrcholů a množina hran⁸, která reprezentuje topologii nějakého systému.

Při modelování v této kapitole budeme ve většině případů chápat vrcholy sítě jako jedince a hrany jako jejich vzájemný vztah, tedy například to, že se jednou za čas setkají. Musíme si ale uvědomit, že při modelování nám nejde o to zjistit, kdy bude jaký člověk nakažen, ale spíše o nalezení obecného chování modelu. Nezáleží tedy na tom, jestli si pod vrcholy představujeme velká letiště a za „setkání“ dvou vrcholů přelet jednoho letadla s mnoha lidmi, nebo vrcholy považujeme za jednotlivé jedince a setkáním myslíme jakoukoliv příležitost předání choroby.

3.1 Klasifikace sítí

Je klíčové si uvědomit, že topologie sítě hraje jednu z nejdůležitějších rolí při šíření choroby. Abychom mohli strukturu sítě zkoumat, zavedme *stupeň vrcholu* k jako počet hran vycházejících z vrcholu k .

Abychom mohli studovat procesy na komplexních sítích, je často potřeba si vytvořit vlastní umělou síť, která bude mít podobné vlastnosti jako nějaká reálná síť. Často užívaný algoritmus pro tvorbu náhodných grafů je založen na tzv. *Erdős–Rényiho modelu* [4, 5], dále jen ER. Tento způsob generování náhodných grafů tvoří sítě, ve kterých mají stupně vrcholů exponenciální rozdělení. Pro grafy vygenerované touto metodou označme p_k pravděpodobnost, že vrchol má stupeň k . Potom platí:

$$p_k = \frac{\langle k \rangle^k}{k!} e^{-\langle k \rangle}, \quad (3.1)$$

kde $\langle k \rangle$ je průměrný stupeň vrcholů v celém grafu.

Ukazuje se však, že stupně vrcholů reálných sítí nemají exponenciální rozdělení, ale spíše mocninné (power law) rozdělení [8, 7]. Proto se pro modelování šíření chorob v poslední době užívá spíše tzv. *scale-free model*, dále jen SF. Pravděpodobnost, že bude mít vrchol stupeň k je pro SF [4, 5]:

$$p_k = Ck^{-\gamma}, \quad (3.2)$$

kde $\gamma > 0$ je parametr určující míru propojenosti a C pouze normalizační konstanta. Normalizační konstanta C zajišťuje, aby $\sum_{k=1}^K p_k = 1$, můžeme ji tedy snadno odvodit. Pro snazší značení budeme užívat běžnou definici Reimannovy funkce zeta [9] $\zeta(s) = \sum_{k=1}^{\infty} k^{-s}$:

$$\sum_{k=1}^{\infty} p_k = C \sum_{k=1}^{\infty} k^{-\gamma} = C\zeta(\gamma). \quad (3.3)$$

Odtud již snadno dostáváme $C = 1/\zeta(\gamma)$, tedy

$$p_k = \frac{k^{-\gamma}}{\zeta(\gamma)}. \quad (3.4)$$

⁸Pod hranou je v tomto případě myšlena dvojice vrcholů. V celém textu budeme uvažovat pouze neorientované hrany, tedy neuspořádané dvojice vrcholů. Více o využití teorie grafů v komplexních sítích se můžeme dočíst v [4, 5]

Rozdílem SF sítí od ER sítí je převážně to, že umožňuje vytvářet vrcholy s velmi vysokým stupněm.

V práci budeme používat také grafy s umělou topologií. Takovéto sítě nemají strukturu ani zdaleka podobnou lidské společnosti, mají však zajímavé vlastnosti, které se mohou hodit pro testování různých hypotéz. Nejprve zavedeme plně propojený graf, tj. graf s n vrcholy ve kterém každý vrchol má stupeň přesně $n - 1$, je tedy propojen se všemi ostatními vrcholy.

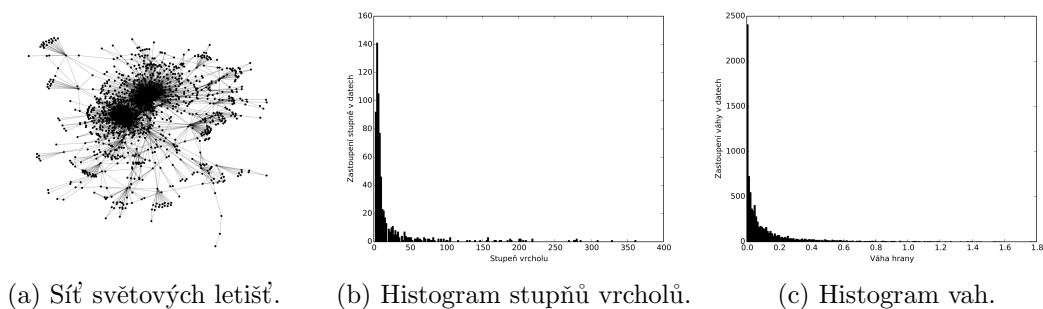
Dále v práci budeme užívat tzv. čtvercovou síť. To je graf ve kterém má každý vrchol stupeň právě čtyři. Je potřeba si uvědomit, že nejde o „tradiční čtverec“, v tom by „okrajové vrcholy“ neměly stupeň čtyři. Každý okrajový vrchol spojíme s jiným okrajovým vrcholem, čímž dosáhneme toho, že všechny vrcholy mají stupeň čtyři. Nyní trochu přesněji. Uvažujme graf $G(V, E)$, kde V je množina vrcholů a E množina hran. Počet vrcholů, tedy velikost množiny V , zvolme tak, aby $|V| = n^2$, kde $n \in \mathbb{N}$. Vrcholy označme postupně čísly $1, \dots, n^2$. Označme E_k množinu hran vrcholu k , poté platí:

$$E_k = \{(k, k - 1 \bmod n^2), (k, k + 1 \bmod n^2), (k, k - n \bmod n^2), (k, k + n \bmod n^2)\}. \quad (3.5)$$

Není nutné naprosto přesně chápat definici tohoto grafu. Stačí si představit čtverec tvořený vrcholy, kde každý vrchol má stupeň čtyři.

3.2 Síť světových letišť

K dispozici jsme měli data obsahující mnoho informací o největších světových letištích. Obsahují přesně 702 letišť a 8152 spojů, tedy hran grafu. Každá hrana je ohodnocena hodnotou, která udává počet letadel, která přeletěla za určitý časový interval. Střední hodnota těchto hran je 0.112. To jsou právě data, která budeme při dalších výpočtech potřebovat. Z histogramu 10(b) si povšimněme, že skutečně



Obrázek 10

většina vrcholů má malý stupeň. I přesto však existují vrcholy s několikanásobně větším stupněm. Takovéto vrcholy jsou označovány jako *hubs* a hrají významnou roli v epidemiologických modelech. Pokud je takovýto *hub*, který je propojen s mnoha dalšími vrcholy, infikován, nemoc se rozšíří mezi mnoha dalších vrcholů. Bude mít mnohem větší vliv na vývoj celého systému než vrchol s průměrným stupněm. Příkladem takovýchto vrcholů v lidské společnosti jsou například prodávači, stevardi, zdravotnický personál, tedy lidé, kteří se setkávají v průběhu dne s mnoha dalšími jedinci.

Odstrašujícím příkladem tohoto efektu může být příběh HIV pozitivního stewarda Gaëtana Dugase [6]. Za svůj život podle všeho vystřídal až 2500 sexuálních partnerů z celého světa, čímž výrazně napomohl šíření AIDS v 80. letech 20. století.

Jak jsme již zmínili, naše data obsahují i váhy jednotlivých hran grafu, jejich histogram je vykreslen na obrázku 10(c). Tyto váhy mají exponenciální rozdělení. Představují počet letadel, která mezi letišti přeletěla za určitý časový interval. Pokud bychom chtěli odhadnout, jak dlouho bude hrana čekat, než po ní proletí letadlo, musíme uvažovat převrácenou hodnotu této hrany. Tedy čím větší bude mít hrana váhu, tím vícekrát po ní za jeden časový interval proletí letadlo a tím kratší čas bude čekat na průlet dalšího letadla. Těchto poznatků využijeme v dalších částech práce.

3.3 Tradiční *SIS* algoritmus na komplexní síti

Jak jsme se snažili ukázat, pro modelování šíření choroby je potřeba pracovat na sítích a brát v úvahu jejich topologii. Nejprve ukažme princip simulace, která je užívána v článku [10]. Poté představíme námi užívanou simulaci, která přináší odlišný pohled na věc. I přes to, že ve výše zmiňovaném článku je popsán algoritmus pro simulace typu *SIR*, budeme popisovat simulaci typu *SIS*, jelikož ji budeme v následujících částech práce užívat.

Uvažujme, že studujeme šíření choroby po neorientovaném grafu $G(V, E)$, kde V je množina vrcholů, tedy jedinců a E množina hran, tedy vztahů dvou jedinců. Jelikož mluvíme o modelu *SIS*, vrcholy se v daném čase mohou nacházet právě v jednom ze stavů S , nebo I . V počátečním čase $t_0 = 0$ je většina vrcholů ve stavu S až na několik málo nakažených jedinců, tedy vrcholů ve stavu I . Pokud je hrana $(i, j) \in E$ složena z jednoho vrcholu ve stavu S a jednoho vrcholu ve stavu I , má vrchol ve stavu S pravděpodobnost α , že se za jednotku času stane infekčním (I). Každý vrchol ve stavu I má zároveň pravděpodobnost β , že se za jednotku času vyléčí, tedy přejde do stavu S . Celý algoritmus poté běží tak, že každou jednotku času vybereme náhodně jeden nakažený vrchol, s pravděpodobností β nastane událost uzdravení \mathcal{I}^- . Pokud tato událost nenastane, vybereme jednu hranu spojující dva libovolné vrcholy S a I a s pravděpodobností α se nakazitelný vrchol stane infekčním. Jednotku času označme jako Δt^9 .

Pro následující algoritmus označme \mathbb{I} množinu nakažených vrcholů, $\mathbb{S} = V \setminus \mathbb{I}$ množinu zdravých vrcholů a $E_{SI} = \{(i, j) \in E; (i \in \mathbb{I} \wedge j \in \mathbb{S}) \vee (i \in \mathbb{S} \wedge j \in \mathbb{I})\}$ množinu hran spojující vrcholy ve stavu S a I . Prvky těchto množin se budou v čase měnit. Je zřejmé, že E_{SI} je jednoznačně určeno množinami \mathbb{I} a E , v algoritmu tedy nemusíme provádět změny přímo množiny E_{SI} , stačí provádět nutné změny v množinách, na kterých je závislý.

⁹V článku [10] je voleno $\Delta t = 1/\beta$.

Data: Graf $G(V, E)$, tj. množina vrcholů V a hran E . Počáteční stav, tj. množina na počátku nakažených vrcholů \mathbb{I}_0 . Parametry α, β . Jednotka času Δt , počáteční čas $t_0 = 0$ a čas t_{max} , kdy chceme simulaci ukončit.

```

1  $\mathbb{I} := \mathbb{I}_0$ ;
2  $t := t_0$ ;
3 while  $t < t_{max}$  do
4    $t := t + \Delta t$ ;
5   náhodně vybereme vrchol  $v \in \mathbb{I}$ ;
6   s pravděpodobností  $\beta$  se uzdraví;
7   if je uzdraven then
8      $\mathbb{I} := \mathbb{I} \setminus \{v\}$ ;
9      $\mathbb{S} := \mathbb{S} \cup \{v\}$ ;
10  else
11    vybereme hranu  $e \in E_{SI}$ , kde  $e = (i, j)$ , jelikož jde o neorientovaný
12    graf, můžeme bez újmy na obecnosti uvažovat  $i \in \mathbb{S} \wedge j \in \mathbb{I}$ ;
13    s pravděpodobností  $\alpha$  se vrchol  $i$  nakazí;
14    if je nakažen then
15       $\mathbb{I} := \mathbb{I} \cup \{i\}$ ;
16       $\mathbb{S} := \mathbb{S} \setminus \{i\}$ ;
17    end
18  end
19 aktualizujeme množinu  $E_{SI}$ ;
20 end

```

Algoritmus 3: Tradiční stochastický algoritmus *SIS* simulace na komplexní síti.

V algoritmu nemáme proměnné, které by nám udávaly zlomek právě nakažených jedinců, což je většinou právě informace, kterou hledáme. Můžeme ji snadno získat tak, že každý časový interval zapíšeme $i := \frac{|\mathbb{I}|}{|V|}$.

Povšimněme si, že tento algoritmus dělá časové kroky konstantní délky a pokaždé dá nejprve možnost na uzdravení a až poté možnost na nakažení libovolně vybraného vrcholu. To se nemusí jevit jako dostatečně odpovídající realitě. Právě ono libovolné vybírání vrcholu, resp. hrany ze sítě dělá snad největší rozdíl od našeho algoritmu popsaného v následující části.

3.4 Nový algoritmus

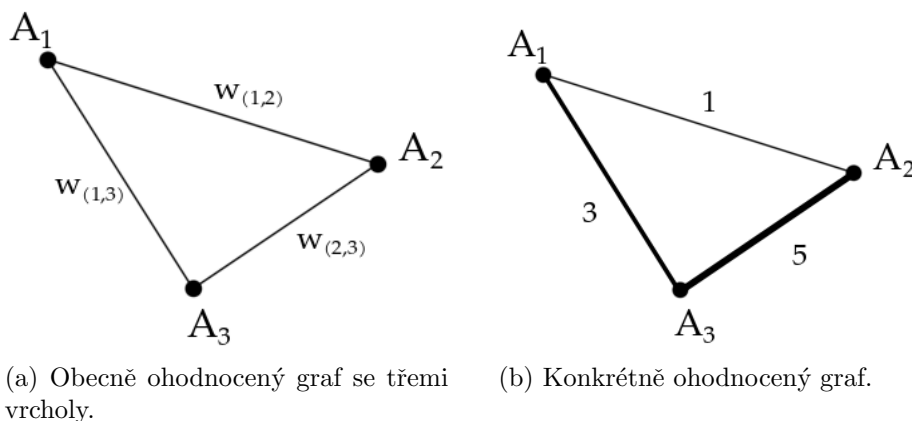
Pro následující algoritmus budeme jako motivaci užívat přenos choroby mezi lidmi jako takovými. Dva lidé, kteří spolu mají nějaký vztah, se jednou za čas setkají a vystaví se tedy možnosti si onemocnění předat. To je hlavní myšlenka, kterou využijeme. Uvažme, že i přes tuto motivaci se nemusí jednat pouze o lidský kontakt. I po síti přenesená data, která mohou nést viry, se pohybují s jakousi pravidelností. Podobně i zprávy o nových událostech se nerozšíří najednou mezi celou společností. Důležité je si uvědomit, že tato setkání, tedy události závislé pouze na topologii sítě se dějí bez jakékoliv závislosti na chorobě¹⁰. Síť bude mít

¹⁰Neuvažujeme katastrofické scénáře epidemií, kdy se lidé straní kontaktu s ostatními, jak jen je možné. Ne všechny choroby jsou tak závažné. Výzkumu šíření chorob tedy více prospěje větší abstrakce a odpoutání se od tolika detailů, které jsou v tomto případě nedůležité.

tuto dynamiku „setkávání“ ať už se v ní šíří choroba, nebo ne. To je nejdůležitější myšlenka oddělující simulaci z předchozí kapitoly od tohoto modelu. Setkání jsou tedy vlastnosti sítě, nikoli choroby.

Uvažujme neorientovaný graf $G(V, E)$, kde V je množina vrcholů, tedy jedinců a E množina hran, tedy vztahů dvojice jedinců. Je ekvivalentní, mluvíme-li o dvou jedincích spojených hranou, nebo o hraně samotné. Za setkání dvou vrcholů spojených hranou považujeme jedinou událost, kdy je možný přenos choroby. Setkat se mohou pouze vrcholy spojené hranou. Pokud je hrana tvořena jedním vrcholem ve stavu S a jedním ve stavu I , k přenosu choroby dojde s pravděpodobností $\alpha \in \langle 0, 1 \rangle$. Pokud je pravděpodobnost $\alpha = 0$, znamená to v podstatě to, že choroba je nepřenositelná a tedy se nerozšíří do společnosti. Pokud je naopak $\alpha = 1$, znamená to, že při každém kontaktu infekčního (I) a nakazitelného (S) se choroba jistě přenesou. Tímto jsme popsali událost a závislosti nakažení při setkání dvou jedinců, tedy vlastnost choroby.

Dále musíme vysvětlit, podle čeho se bude řídit frekvence samotných setkání. Pro každou hranu chceme najít její *waiting time*, tedy čas, který budeme muset čekat než po hraně dojde k dalšímu setkání. Tato veličina závisí pouze na síti a má exponenciální rozdělení. Parametry pro *waiting time* představují právě váhy hran sítě. To jsou četnosti setkání, tedy počty setkání za určitý časový interval. Čím větší je hodnota hrany, tím častěji za určitý časový úsek se vrcholy hrany setkají a tím kratší dobu budou muset čekat na další setkání.



Obrázek 11: Ohodnocená komplexní síť.

Uvažujme graf zobrazený v 11. Hrana (A_1, A_2) má hodnotu $w_{1,2} = 1$. To znamená, že za jeden časový interval Δt se vrcholy A_1, A_2 setkají průměrně jednou. Na rozdíl od toho hrana (A_2, A_3) má hodnotu $w_{2,3} = 5$. To znamená, že za jeden časový interval Δt se vrcholy A_2, A_3 setkají průměrně pětkrát. Střední hodnota času, který musí vrcholy A_1, A_2 počkat na setkání tedy bude $\langle t_{(1,2)} \rangle = \frac{\Delta t}{1}$. Tato střední hodnota bude pro vrcholy A_2, A_3 pětkrát menší $\langle t_{(2,3)} \rangle = \frac{\Delta t}{5}$. Je nezbytné si uvědomit, že časový krok Δt je v tomto případě pouhý parametr, kterým můžeme čas škálovat na reálné jednotky. Na rozdíl od algoritmu z předchozí kapitoly nijak neovlivňuje samotnou simulaci.

Vidíme, že *waiting time* je nepřímo úměrný četnosti setkání, tedy váze hrany. Střední hodnota čekacích časů hrany (i, j) tedy bude mít hodnotu $\langle t_{(i,j)} \rangle = \frac{1}{w_{(i,j)}}$. Víme také, že tyto časy mají exponenciální rozdělení. Z toho plyne, že parametr

exponenciálního rozdělení¹¹ bude $\beta = \langle x \rangle = \frac{1}{w_{(i,j)}}$. Hodnoty pro *waiting time* hrany (i, j) s váhou $w_{(i,j)}$ tedy budeme vybírat z $\text{Exp}(\frac{1}{w_{(i,j)}})$.

Proces uzdravení nijak nezáleží na topologii sítě. Nakažený jedinec se v modelech typu *SIS* uzdraví, ať už je to jediný vrchol celého grafu, nebo součást velké komplexní sítě. Tento proces vůbec nezáleží na topologii sítě, pouze na parametrech choroby. Z úvodních kapitol o stochastických modelech můžeme vyvodit, že čas, který musí jedinec čekat na uzdravení, bude také exponenciálně rozložen. Toto exponenciální rozdělení však nebude záviset na síti, bude záviset pouze na chorobě samotné. Střední hodnotu a tedy i parametr tohoto exponenciálního rozdělení označme β_I . Tento parametr má jednotky času v základním tvaru. Obdobně by probíhal proces přechodu ze stavu R do stavu S .

Celá simulace poté probíhá tak, že na počátku vylosujeme časy na setkání pro všechny hrany (i pro $S - S$ hrany) i časy na uzdravení všem na počátku infekčním vrcholům. Všechny tyto časy seřadíme v seznamu tak, jak půjdou po sobě. Jakmile dojde k času, kdy má nastat setkání na hraně, ověříme, zda je ona hrana prvkem množiny E_{SI} ¹². Pokud ano, s pravděpodobností α se nakazitelný vrchol hrany stane infekčním. Aktualizujeme množinu \mathbb{I} , vylosujeme hraně čas¹³ do nového setkání a postupujeme k další události v našem uspořádaném seznamu dějů. Pokud v seznamu narazíme na událost, kdy se má vrchol uzdravit, uzdravíme ho¹⁴ a aktualizujeme množinu \mathbb{I} .

Ponechejme si značení z předchozí kapitoly a přidejme k němu uspořádanou množinu \mathbb{T} , která bude označovat náš uspořádaný seznam událostí. Každý prvek, tedy událost tohoto seznamu, musí obsahovat čas kdy nastane a místo kde nastane, tj. nějaký vrchol/hrana. Tím, že určíme zda jde o událost na vrcholu, nebo hraně, určíme také zda půjde o uzdravení, nebo o pokus o nakažení.

¹¹Více o exponenciálním rozdělení a našem značení v příloze.

¹²Značení užíváme stejné jako v předchozí kapitole.

¹³To je nutné provést i pokud s hranou nic neprovedeme, tj. pokud hrana není složena z jednoho vrcholu S a jednoho vrcholu I .

¹⁴Uzdravíme ho jistě. Tato událost již nezávisí na žádném dalším parametru. Ona kýžená stochasticita je dána již výběrem času uzdravení z exponenciálního rozdělení s parametrem β_I .

Data: Graf $G(V, E)$, tj. množina vrcholů V a hran E s jednotlivými váhami $w_{(i,j)}$. Počáteční stav, tj. množina na počátku nakažených vrcholů \mathbb{I}_0 . Parametry α, β_I . Počáteční čas $t_0 = 0$ a čas t_{max} , kdy chceme simulaci ukončit.

```

1  $\mathbb{I} := \mathbb{I}_0$ ;
2  $t := t_0$ ;
3  $\mathbb{T}$  je prázdný;
4 pro všechny hrany  $(i, j) \in E$  vylosujeme čas  $t_{(i,j)} \sim \text{Exp}(\frac{1}{w_{(i,j)}})$  a zařadíme
   do  $\mathbb{T}$ ;
5 pro všechny vrcholy  $v \in \mathbb{I}_0$  vylosujeme čas  $t_v \sim \text{Exp}(\beta_I)$  a zařadíme do  $\mathbb{T}$ 
   while  $t < t_{max}$  do
6     vezmeme první prvek  $k$  z  $\mathbb{T}$ ;
7      $t := t + t_k$ ;
8     if  $k$  je hrana then
9         if  $k \in E_{SI}$ , kde  $i \in \mathbb{I}, j \in \mathbb{S}$  then
10            s pravděpodobností  $\alpha$  se vrchol  $j$  nakazí;
11            if  $j$  se nakazí then
12                 $\mathbb{I} := \mathbb{I} \cup \{j\}$ ;
13                 $\mathbb{S} := \mathbb{S} \setminus \{j\}$ ;
14                vylosujeme  $t_j \sim \text{Exp}(\beta_I)$  a do  $\mathbb{T}$  zařadíme  $t_j + t$ ;
15            end
16        end
17        pro  $k = (i, j)$  vylosujeme nové  $t_k \sim \text{Exp}(\frac{1}{w_k})$  a do  $\mathbb{T}$  zařadíme  $t_k + t$ ;
18    else // jde tedy o vrchol, který má být uzdraven
19         $\mathbb{I} := \mathbb{I} \setminus \{k\}$ ;
20         $\mathbb{S} := \mathbb{S} \cup \{k\}$ ;
21    end
22    aktualizujeme množinu  $E_{SI}$ ;
23    odstraníme první prvek z  $\mathbb{T}$ ;
24 end

```

Algoritmus 4: SIS algoritmus oddělující vlastnosti sítě.

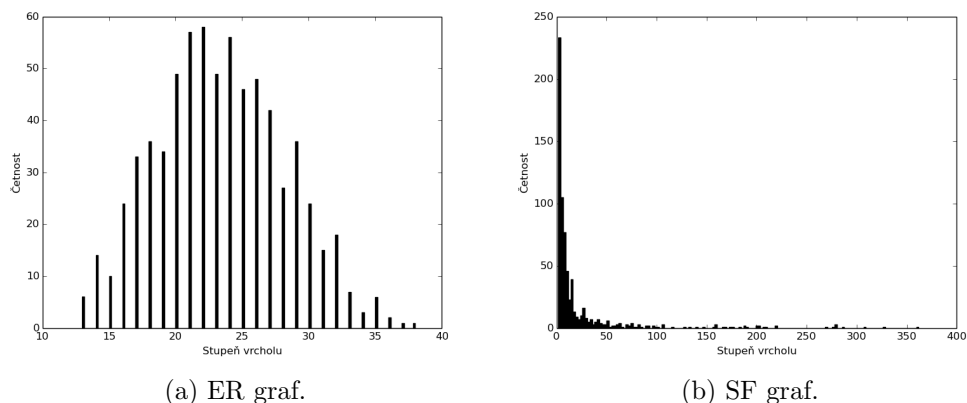
Jelikož tento algoritmus užívá naprosto odlišnou parametrizaci nemůžeme jeho výsledky přímo porovnat s algoritmem z předchozí kapitoly.

Povšimněme si však, že tento algoritmus užívá velmi snadno interpretovatelné parametry. Parametr nemoci β_I není nic jiného než střední hodnota délky trvání nemoci. Hodnota α , tedy pravděpodobnost nakažení při setkání, se dá jistě snadno zjistit se základními znalostmi o přenosu zkoumané choroby. Jednotlivé váhy hran jsou četnosti „setkání“. Stejně tak to mohou být četnosti přeletů letadel mezi letišti, které jsou velmi snadno zjistitelné.

4 Aplikace algoritmu

Abychom ukázali výrazný podíl topologie sítě na šíření nemoci, numericky jsme simulovali vývoj na třech různých typech sítí. Užili jsme umělou síť čtverců, tedy síť, kde každý vrchol má stupeň právě čtyři. Dále jsme užili síť světových letišť jako příklad *scale-free* topologie a síť vytvořenou pomocí ER modelu tak, aby měla stejný počet vrcholů a přibližně stejný počet hran jako SF.

Konkrétně se tedy jedná o SF graf s počtem vrcholů $N_{SF} = 702$, počtem hran $|E_{SF}| = 8152$ a ER graf s počtem vrcholů $N_{ER} = 702$, počtem hran $|E_{ER}| = 8185$. Jelikož počet vrcholů čtvercového grafu může být pouze druhá mocnina přirozeného čísla, pro čtverce volíme počet vrcholů $N' = 676 = 26^2$. Počet hran je tedy $|E'| = 2N' = 1352$.



Obrázek 12: Histogram stupňů vrcholů.

Porovnáváme-li histogramy 12, vidíme, že rozložení stupňů v ER grafu má v porovnání s SF sítí malý rozptyl. To znamená, že stupně jednotlivých vrcholů v ER budou poměrně podobné. Na rozdíl od toho v SF je rozptyl několikanásobně větší. Můžeme si také povšimnout samotných hodnot stupňů. V SF je mnoho vrcholů s velmi nízkým stupněm. Na takovéto vrcholy bude choroba jen velmi těžko dosahovat. Stejně tak jsou tady však vrcholy s extrémně vysokým stupněm, které v případě, že jsou infekční, zapříčiní velmi rychlý start epidemie. Jelikož mají vysoký stupeň, tak se velmi snadno infekčními stanou.

Počáteční počet infekčních byl $I_0 = 20$ pro SF a ER. Abychom mohli simulace na všech grafech porovnávat, zvolili jsme pro čtverce $I'_0 = 19$, čímž jsme alespoň přibližně zachovali zlomek na počátku nakažených $i_0 = \frac{I_0}{N_{SF}} = \frac{I_0}{N_{ER}} \simeq \frac{I'_0}{N'}$. Pro každý běh simulace jsme znova náhodně vybrali množinu na počátku infekčních vrcholů.

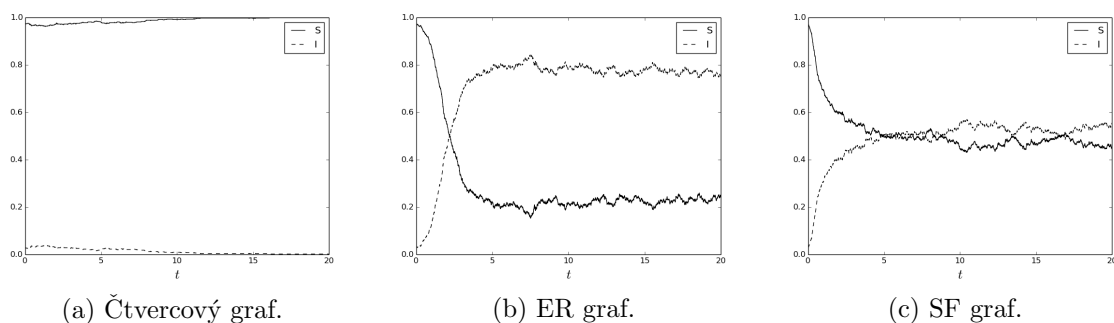
Nejprve jsme volili konstantní váhu hran $w = 1$, později jsme je vybírali z exponenciálního rozdělení¹⁵. Střední hodnotu a tedy i parametr exponenciálního rozdělení pro dobu trvání nemoci jsme vybrali $\beta_I = 2$.

Parametr nemoci α jsme volili postupně $\alpha = 0.1, 0.2, \dots, 1.0$. Pro každý graf a každou hodnotu parametru jsme simulaci prováděli právě desetkrát, abychom předešli tvorbě chybných závěrů z extrémních a ojedinělých případů, které při stochastickém modelování mohou nastat.

4.1 Simulace s konstantními vahami

Nejprve jsme simulace prováděli na zmíněných sítích s konstantními vahami hran. Můžeme tak snadno pozorovat velký vliv topologie sítě nijak nezměněné ohodnocením. Simulace jsme prováděli do času $t_{max} = 20$. Na grafech 13 je na vodorovné ose čas a na svislé ose zlomek zdravých, resp. infekčních jedinců.

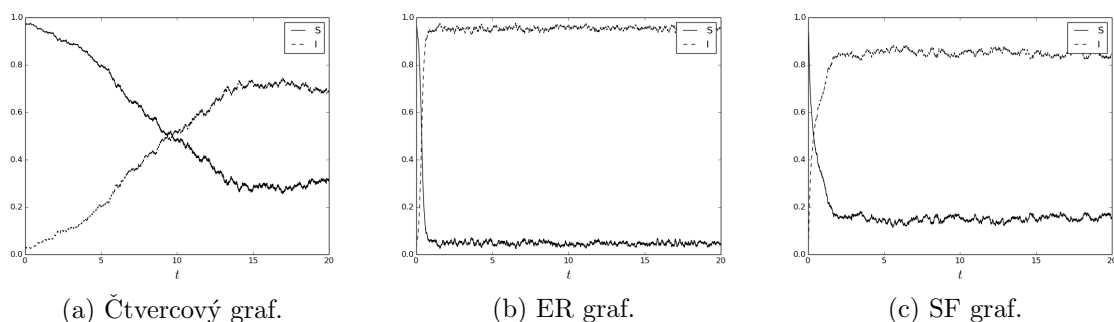
¹⁵Proč právě z exponenciálního rozdělení je objasněno v kapitole věnované datům o letištích.



Obrázek 13: Simulace pro $\alpha = 0.1$.

Na grafu 13(a) můžeme vidět, že pro malé $\alpha = 0.1$ choroba brzy vymírá¹⁶. To je způsobeno velmi malou konektivitou v grafu. Jedinec má možnost setkat se pouze se čtyřmi jinými vrcholy.

Porovnáme-li graf 13(b) s grafem 13(c) vidíme, že pro stejné hodnoty je výsledné ekvilibrium zlomku infekčních pro ER síť vyšší než pro SF síť. To je způsobeno těžce dosažitelnými vrcholy s nízkým stupněm, které se ve velkém počtu nacházejí v SF sítích¹⁷. Vrcholy s extrémně vysokým stupněm v SF grafu na rozdíl od toho způsobují, že se choroba rozšíří rychleji než v ER. To můžeme pozorovat v počátečních fázích vývoje systému v 13(b), 13(c).



Obrázek 14: Simulace pro $\alpha = 0.5$.

V simulaci s $\alpha = 0.5$ je vidět, že na čtvercovém grafu je choroba již tak přenositelná¹⁸, že i na velmi řídkce propojené síti je schopna se rozšířit a nakazit dokonce více než polovinu společnosti. Pro takto zvolené α je již velmi nepravděpodobné, že by nemoc vymřela a vymizela ze společnosti.

Simulace na ER síti 14(b) ukazuje, že v takovéto topologii stačí $\alpha = 0.5$, aby se nemoc rozšířila do naprosté většiny společnosti. Počet zdravých jedinců je minimální. V SF síti 14(c) se nemoc opět rozšíří v počátečních fázích vývoje systému velmi rychle, těžko dosažitelné vrcholy však stále zůstávají nenakažené, což výrazně snižuje zlomek infekčních jedinců.

¹⁶Tím je myšleno, že se poslední infekční jedinec uzdravil dříve než chorobu stihl předat. Tím se nemoc již nevyskytuje nikde v systému.

¹⁷To ostatně můžeme vidět z histogramu 12(b).

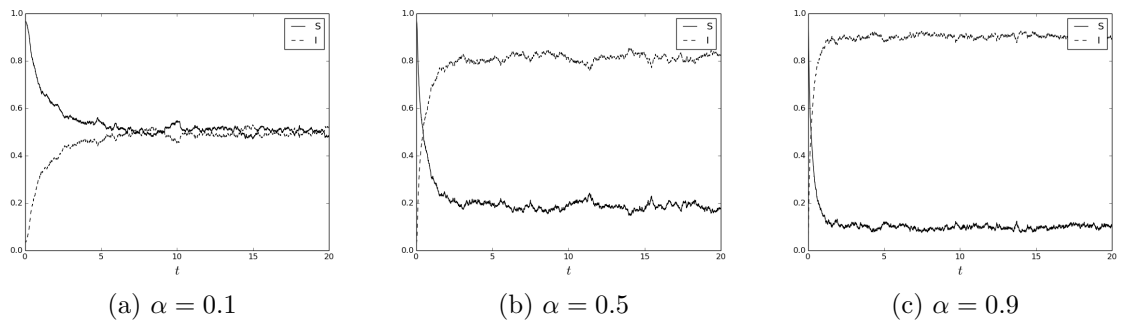
¹⁸Právě parametr α nám udává míru přenositelnosti nemoci.

4.2 Rozložení vah a simulace na ohodnocené síti

Než přejdeme k simulacím na ohodnocených grafech, je nutné si rozmyslet, jak přiřazovat hranám váhy. Z kapitoly o letištích a konkrétně z histogramu 10(c) je vidět, že rozložení samotných vah můžeme aproximovat pomocí exponenciálního rozdělení.

Jelikož existuje korelace mezi vahou hrany a průměru stupňů jejích vrcholů, nemůžeme přiřazovat váhy z exponenciálního rozdělení libovolným hranám. Abychom ukázali tento podstatný rozdíl, porovnáme simulace s náhodně rozloženými vahami a simulace s konkrétními váhami z dat.

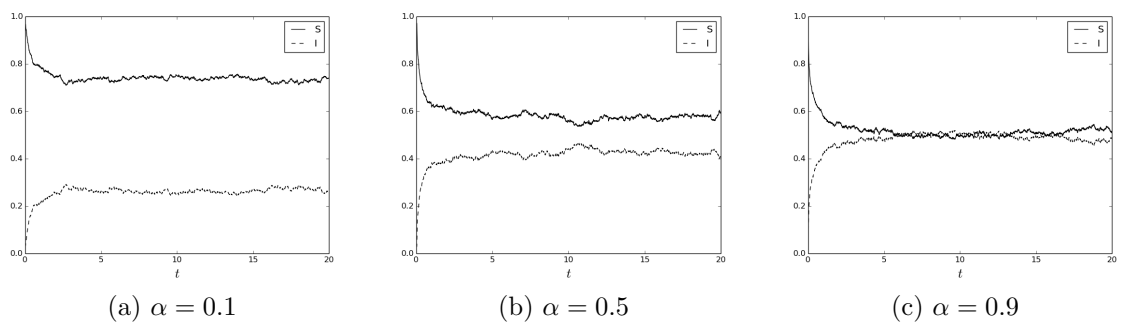
Nejprve provedeme simulace s náhodným přiřazením vah. Vybereme střední hodnotu exponenciálního rozdělení, kolem které chceme mít rozvržené váhy. Dále již jen náhodně z takového rozdělení vybíráme váhy jednotlivým hranám. Aplikujeme-li tento postup na SF síti se střední hodnotou vah $\langle w \rangle = 1$ dostáváme výsledky velmi podobné těm v předchozí kapitole.



Obrázek 15: Simulace na SF síti s náhodně rozdanými váhami hran.

Jelikož naše data o letištích mají ohodnocené hrany, můžeme výsledky simulace, která náhodně rozděljuje váhy, porovnat se simulací na skutečných vahách.

Musíme si uvědomit, že abychom došli k porovnatelné simulaci, musíme škálovat váhy z reálných dat tak, aby měly stejnou střední hodnotu, jakou jsme užili v předchozím případě, tj. $\langle w \rangle = 1$. Pokud bychom tak neučinili, museli bychom změnit střední hodnotu trvání nemoci. Je jedno kterou veličinu budeme škálovat, konečný důsledek bude jen jiná jednotka času.

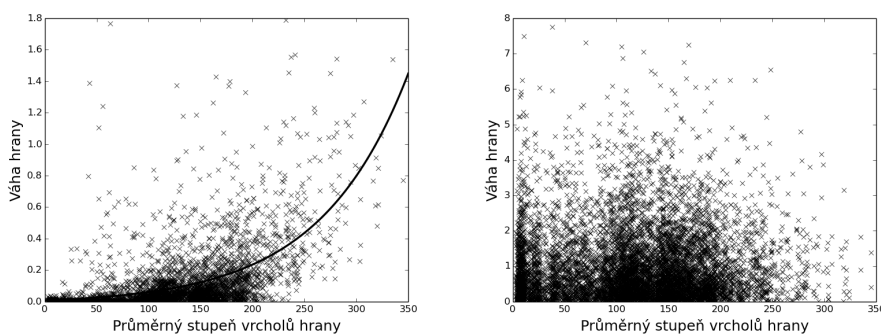


Obrázek 16: Simulace na SF síti s reálnými váhami hran.

Pro takto upravená data již můžeme provést stejnou simulaci jako v předchozím případě.

Je naprosto zřejmé, že průběh na reálných datech se kvalitativně liší od průběhu na síti s náhodně zvolenými váhami hran. Vidět jsou oba dva důležité efekty SF sítí. Prvním z nich je rychlé rozšíření nemoci v počátečním stadiu, které je v případě na grafu 16 ještě podtrženo tím, že vrcholy s vysokým stupněm mají mnoho hran s vysokými váhami. Druhým efektem je to, že se nemoc tak snadno nerozšíří do naprosté většiny společnosti. V případě grafu 15 je to způsobeno tím, že existují těžko dosažitelné vrcholy s velmi malým stupněm. Podíváme-li se na simulaci na reálných datech na grafu 16, je tento efekt podtržen ještě tím, že tyto izolované vrcholy mají velmi nízké váhy hran.

Předchozí srovnání nás vede k domněnce, že váhy a stupně vrcholů spolu nějakým způsobem korelují. Abychom ukázali závislost váhy hrany na stupních jejích vrcholů vykreslíme graf 17(a) pro skutečné rozložení vah z dat o letištích a graf 17(b) s rozložením vah v síti vygenerovaném náhodně při jednom z běhů simulace. Na svislé ose je váha jednotlivých hran a na vodorovné ose aritmetický průměr příslušných vrcholů konkrétní hrany.



(a) Data o letištích s nabitou exponenciální funkcí. (b) Rozložení vah na grafu při jedné ze simulací.

Obrázek 17: Závislost váhy hrany na průměrném stupni jejích vrcholů.

Je vidět, že v případě reálného rozložení vah z dat o letištích 17(a), hrany s nízkou vahou z pravidla spojují vrcholy s nízkým stupněm. Zvyšujeme-li váhu hrany, zvyšuje se i její průměrný stupeň. S touto změnou se však pojí i značný růst rozptylu. K ověření korelace jsme využili Spearmanův korelační koeficient¹⁹ [11] a jednu ze statistických funkcí balíčku *scipy* programovacího jazyka *Python*. Korelační koeficient reálného případu je $\rho_1 \approx 0.640$ a p-hodnota $p_1 \approx 0.00$, což ukazuje, že hodnoty spolu skutečně korelují.

Naproti tomu, v případě, kdy jsme váhy přiřazovali hranám naprosto náhodně žádná taková korelace nevzniká 17(b). Korelační koeficient je v tomto případě $\rho_2 \approx 0.007$ a p-hodnota $p_2 \approx 0.54$, což značí, že data spolu skutečně nekorelují.

Je tedy zřejmé, že velmi záleží na váze jednotlivých hran sítě. Nový algoritmus tuto vlastnost v úvahu bere. Na rozdíl od toho [10] a jemu podobné články tuto vlastnost nijak nevyšetřují.

¹⁹Více o Spearmanově korelačním koeficientu v příloze.

5 Závěr

V první části práce jsme shrnuli základy výzkumu šíření epidemií ve společnosti. Dále jsme vytvořili způsob simulace, který odděluje vlastnosti choroby od vlastností topologie společnosti. Algoritmus vychází přímo z dějů v lidské společnosti, jeho parametry jsou snadno zjistitelné a interpretovatelné. Pomocí něj jsme ukázali nutnost některých vlastností, které tradičně užívané algoritmy neuvažují.

Dalším námětem ke studiu je jistě analytický pohled na nový algoritmus. Výzvou může být hledání aproximovaných časových řešení systému, nebo například hraničních parametrů, které způsobují vymizení choroby.

Reference

- [1] PEREZ, Fernando a Brian E. GRANGER. IPython: A System for Interactive Scientific Computing. *Computing in Science*. 2007, **9**(3), 21-29. DOI: 10.1109/MCSE.2007.53. ISSN 1521-9615. Dostupné také z: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4160251>
- [2] Epidemic model. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2001-, 2015-12-09 [cit. 2016-01-02]. Dostupné z: https://en.wikipedia.org/wiki/Epidemic_model
- [3] E (mathematical constant). In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2016-02-09]. Dostupné z: [https://en.wikipedia.org/wiki/E_\(mathematical_constant\)](https://en.wikipedia.org/wiki/E_(mathematical_constant))
- [4] NEWMAN, M. *Networks: an introduction*. Oxford: Oxford University Press, 2010, xi, 772 s. ISBN 978-0-19-920665-0.
- [5] HAVLIN, Shlomo a Reuven COHEN. *Complex networks structure, robustness, and function*. New York: Cambridge University Press, 2010. ISBN 978-051-1776-618.
- [6] BARABASI, Albert-Laszlo. *Linked the new science of networks*. [Online-Ausg.]. New York: Basic Books, 2002. ISBN 978-078-6746-965.
- [7] PASTOR-SATORRAS, Romualdo a Alessandro VESPIGNANI. Epidemic Spreading in Scale-Free Networks. *Physical Review Letters*. 2001, **86**(14), 3200-3203. DOI: 10.1103/PhysRevLett.86.3200. ISSN 0031-9007. Dostupné také z: <http://link.aps.org/doi/10.1103/PhysRevLett.86.3200>
- [8] PASTOR-SATORRAS, Romualdo a Alessandro VESPIGNANI. Epidemic dynamics and endemic states in complex networks. *Physical Review E*. 2001, **63**(6). DOI: 10.1103/PhysRevE.63.066117. ISSN 1063-651x. Dostupné také z: <http://link.aps.org/doi/10.1103/PhysRevE.63.066117>
- [9] Riemann zeta function. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2016-02-13]. Dostupné z: https://en.wikipedia.org/wiki/Riemann_zeta_function
- [10] HOLME, Petter a Taro TAKAGUCHI. Time evolution of predictability of epidemics on networks. *Physical Review E*. American Physical Society, 2015, **91**(4). DOI: 10.1103/PhysRevE.91.042811. ISSN 1539-3755. Dostupné také z: <http://link.aps.org/doi/10.1103/PhysRevE.91.042811>
- [11] Spearman's rank correlation coefficient. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2016-03-08]. Dostupné z: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

Přílohy

Metoda nejmenších čtverců

Metoda nejmenších čtverců je statistická metoda pro nalezení vhodných parametrů θ funkce $h_\theta(\mathbf{x})$, pro které funkce $h_\theta(\mathbf{x})$ nejvíce vyhovuje naměřeným datům, tj. uspořádaným dvojicím (\mathbf{x}, \mathbf{y}) ²⁰.

Uvažujme případ, kdy jsme naměřili m dvojic (\mathbf{x}, \mathbf{y}) a známe funkci $\mathbf{y} = h_\theta(\mathbf{x})$, které by hodnoty měly odpovídat. Zavádíme tzv. *cost function* $J(\theta)$:

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \|h_\theta(\mathbf{x}_i) - \mathbf{y}_i\|^2,$$

jejíž hodnoty představují cosi jako míru chyby, které jsme se dopustili volbou určitých parametrů θ . Nejvhodnější parametry funkce $h_\theta(\mathbf{x})$ budou tedy taková θ_{opt} , pro která:

$$\min_{\theta} J(\theta) = J(\theta_{opt}).$$

Pro reálný výpočet vhodných parametrů užíváme v práci optimalizované balíčky *scipy* programovacího jazyka IPython [1], které jsou postaveny právě na této metodě.

Exponenciální rozdělení

Exponenciální rozdělení je spojitě pravděpodobnostní rozdělení, jehož hustota pravděpodobnosti je pro $x > 0$ dána rovnicí:

$$f(x) = \lambda e^{-\lambda x},$$

kde $\lambda \in \mathbb{R}$ je parametr rozdělení. My budeme v celé práci užívat jinou parametrizaci, kde budeme užívat parametr $\beta = \lambda^{-1}$, tedy:

$$f(x) = \frac{1}{\beta} e^{-\frac{1}{\beta} x}.$$

Střední hodnota tohoto rozdělení je:

$$\langle x \rangle = \frac{1}{\lambda} = \beta.$$

Rozptyl:

$$\text{Var}(x) = \frac{1}{\lambda^2} = \beta^2.$$

Skutečnost, že jsme vybrali hodnotu t z exponenciálního rozdělení s parametrem $\beta = \langle x \rangle$ budeme značit:

$$t \sim \text{Exp}(\beta).$$

²⁰Naměřená data mohou být, stejně jako parametry θ , celé vektory hodnot.

Spearmanův korelační koeficient

Pomocí Spearmanova korelačního koeficientu [11] můžeme měřit závislost dvou veličin pomocí pořadí obou veličin. Říká nám, jak je silná přímá (nepřímá) úměrnost mezi pořadími obou statistických znaků. Většinou je značen ρ , nebo r_s . Nabývá hodnot z intervalu $\langle -1, 1 \rangle$. Pokud je hodnota $\rho = 1$, nebo $\rho = -1$, jsou pořadí veličin shodná, nebo přesně opačná. Hodnota $\rho = 0$ nám na rozdíl od toho říká, že hodnoty obou veličin spolu v tomto smyslu vůbec nesouvisí.

Uvažujme dvojice dat, uspořádané dvojice $(x_1, y_1), \dots, (x_n, y_n)$. Tyto hodnoty uspořádáme podle velikosti a seřazená data označíme $(r_{x_1}, r_{y_1}), \dots, (r_{x_n}, r_{y_n})$. Spearmanův korelační koeficient poté vypočteme jako standardní (Pearsonův) korelační koeficient z pořadí obou veličin, jež je možné upravit do tvaru:

$$\rho = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^n (r_{x_i} - r_{y_i})^2.$$