**UNICORN VYSOKÁ ŠKOLA S.R.O.**

BAKALÁŘSKÁ

PRÁCE

**2023 Tomáš JIRKŮ**

**UNICORN VYSOKÁ ŠKOLA S.R.O.**

**Softwarový vývoj**



**BAKALÁŘSKÁ PRÁCE**

**Modelování šíření řetězových emailů**

**Autor BP:** Tomáš Jirků

**Vedoucí BP:** RNDr. Mgr. Horáček Jaroslav, Ph.D.

Obsah obrázku text, Písmo, Grafika, design

Popis byl vytvořen automaticky

**Katedra Informačních technologií** **Akademický rok 2022/2023**

**ZADÁNÍ BAKALÁŘSKÉ PRÁCE**

**Jméno a příjmení**



**Tomáš Jirků**



**Studijní program**

**Softwarový vývoj**

**Název práce**

**Modelování šíření řetězových emailů**

**Cíl**

**Bakalářská práce se věnuje simulování šíření řetězových emailů v populaci. Cílem práce je pomocí poznatků z teorie sítí, studií chování populace na internetu a agentového modelování rozšířit existující modely šíření řetězových emailů. Součástí práce je vytvoření nástroje pro načítání sítí na základě reálných dat a vizualizaci šíření řetězových emailů v těchto sítích. S využitím tohoto nástroje bude provedena analýza a diskuze vhodnosti a přínosu stávajících i vlastních modelů.**

**Osnova**

**Úvod**

**1. Teoretická část**

**1.1 Existující modely šíření v sítích**

**1.2 Analýza vhodnosti použití existujících modelů šíření**

**1.3 Rozšíření existujících modelů**

**2. Praktická část**

**2.1 Modelování prostředí emailové sítě**

**2.2 Aplikace existujících modelů šíření na simulované prostředí**

**2.3 Aplikace upravených modelů šíření na simulované prostředí**

**2.4 Analýza výsledků a jejich porovnání**

**Závěr**

**Doporučená literatura**

**Barabási, A.-L., & Pósfai, M. (2016). Network science. Cambridge University Press.**

**Gürbüz, B., Mawengkang, H., Husein, I., & Weber, G.-W. (2021). Rumour propagation: An operational research approach by computational and information theory. Central European Journal of Operations Research. https://doi.org/10.1007/s10100-020-00727-0**

Obsah obrázku text, emblém, symbol, logo

Popis byl vytvořen automaticky

|  |  |  |
| --- | --- | --- |
| **Unicorn Vysoká škola s.r.o.** | **Tel: (+420) 221 400 886** | **IČO: 27169511** |
| **V Kapslovně 2767/2** | **info@unicornuniversity.net** | **DIČ: CZ699004029** |
| **CZ-130 00 Praha 3** | **www.unicornuniversity.net** | **č.ú.: 35-2147930287/0100** |
|  |  | **KB – Praha 1, pobočka 42** |

Obsah obrázku text, Písmo, logo, Grafika

Popis byl vytvořen automaticky

**Iribarren, J. L. (2007). Information diffusion epidemics in social networks. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.1464631**

**Kopecký, K., Kožíšek, M., Szotkowski, R., & Kasáčková, J. (2018). Starci Na Netu (Výzkumná Zpráva).**

**Liben-Nowell, D., & Kleinberg, J. (2008). Tracing information flow on a global scale using internet chain-letter data. Proceedings of the National Academy of Sciences, 105(12), 4633–4638. https://doi.org/10.1073/pnas.0708471105**

**Nekovee, M., Moreno, Y., Bianconi, G., & Marsili, M. (2007). Theory of rumour spreading in complex social networks. Physica A: Statistical Mechanics and Its Applications, 374(1), 457–470. https://doi.org/10.1016/j.physa.2006.07.017**

**Newman, M. E., Forrest, S., & Balthrop, J. (2002). Email networks and the spread of computer viruses. Physical Review E, 66(3). https://doi.org/10.1103/physreve.66.035101**

**Vedoucí bakalářské práce**

**RNDr. Mgr. Jaroslav Horáček, Ph.D.**



**Adresa pracoviště**

 **V Kapslovně 2767/2, 13000 Praha 3**

**Datum zadání bakalářské práce**

**28.01.2022**



**Termín odevzdání bakalářské práce**

 **Podle rozhodnutí rektora**

**V Praze dne 11.04.2023**

**Obsah obrázku rukopis, skica, Dětské kresby, kaligrafie

Popis byl vytvořen automaticky**

**.............................................................**

**doc. Ing. Jan Čadil, Ph.D.**

**Rektor**

**Obsah obrázku text, emblém, symbol, logo

Popis byl vytvořen automaticky**

|  |  |  |
| --- | --- | --- |
| **Unicorn Vysoká škola s.r.o.** | **Tel: (+420) 221 400 886** | **IČO: 27169511** |
| **V Kapslovně 2767/2** | **info@unicornuniversity.net** | **DIČ: CZ699004029** |
| **CZ-130 00 Praha 3** | **www.unicornuniversity.net** | **č.ú.: 35-2147930287/0100** |
|  |  | **KB – Praha 1, pobočka 42** |

**Čestné prohlášení**

Prohlašuji, že jsem svou bakalářskou práci na téma Modelování šíření řetězových emailů vypracoval samostatně pod vedením vedoucího bakalářské práce a s použitím výhradně odborné literatury a dalších informačních zdrojů, které jsou v práci všechny citovány a jsou také uvedeny v seznamu použitých zdrojů.

Jako autor této bakalářské práce dále prohlašuji, že v souvislosti s jejím vytvořením jsem neporušil autorská práva třetích osob a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb.

Dále prohlašuji, že odevzdaná tištěná verze bakalářské práce je shodná s verzí, která byla odevzdána elektronicky.

V……………………. dne ……….. …….……………………………

Tomáš Jirků

*Vzor:* ***PODĚKOVÁNÍ*** *vedoucímu BP, konzultantům, odborníků, spolupracovníkům za poskytnuté rady a podkladové materiály apod.) –* ***není povinné***

**Poděkování**

Např: Děkuji vedoucímu bakalářské práce Jméno Příjmení (i s tituly) za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé bakalářské práce…



**Modelování šíření řetězových emailů**

**Modeling the spread of chain emails**

*Vzor: ABSTRAKT A KLÍČOVÁ SLOVA*

**Abstrakt**

Abstrakt česky. Abstrakt krátce a výstižně charakterizuje obsah závěrečné práce. Zpravidla obsahuje informace o stanovených cílech, použitých metodách, postupu řešení a výsledcích výzkumu. Může obsahovat krátkou informaci o použitých zdrojích. Délka abstraktu je zpravidla 100–500 slov.

Klíčová slova: klíčová slova práce, minimálně 5, maximálně 10

**Abstract**

Zde umístěte překlad abstraktu do anglického jazyka. Česky a anglicky psané abstrakty musí být totožné. Student/ka zodpovídá za jazykovou správnost anglického překladu. V případě, že se anglická a česká verze nevejdou na jednu stránku, umístěte celý překlad na samostatnou stránku.

Keywords: klíčová slova v anglickém jazyce

*Vzor:* ***OBSAH*** *– hierarchické uspořádání číslovaných názvů kapitol a podkapitol, včetně všech příloh, spolu s čísly jejich stran. Dále se uvádí Seznam obrázků/tabulek/grafů. Pozn.: počet a názvy kapitol samozřejmě odpovídají charakteru konkrétní práce.*

**Obsah**

[Úvod 9](#_Toc55215785)

[1 Teoretická část (není název kapitoly) 10](#_Toc55215786)

[1.1 Nadpis úrovně 2 10](#_Toc55215787)

[1.1.1 Nadpis úrovně 3 10](#_Toc55215788)

[1.1.2 Nadpis úrovně 3 10](#_Toc55215789)

[1.2 Nadpis úrovně 2 10](#_Toc55215790)

[1.2.1 Nadpis úrovně 3 10](#_Toc55215791)

[1.2.2 Nadpis úrovně 3 10](#_Toc55215792)

[2 Praktická část/Empirická část/Vlastní práce (není název kapitoly) 11](#_Toc55215793)

[2.1 Nadpis úrovně 2 11](#_Toc55215794)

[Závěr 13](#_Toc55215795)

[Seznam použitých zdrojů 14](#_Toc55215796)

[Seznam obrázků (existují-li) 15](#_Toc55215797)

[Seznam grafů (existují-li) 17](#_Toc55215798)

[Seznam příloh (existují-li) 18](#_Toc55215799)

[Příloha A – Název přílohy 19](#_Toc55215800)

[Příloha B – Název přílohy 20](#_Toc55215801)

*Vzor:* ***ÚVOD*** *(cca 5-10 normostran)*

# Úvod

# Řetězové emaily a jejich modely šíření

## Definice a původ řetězového emailu

Řetězový email se svým obsahem snaží přimět příjemce k jeho přeposlání k co nejvíce dalším uživatelům. Obsahem často bývají nepravdivé informace, různé podvody, či různé manipulativní zprávy útočící na emoce. [7] Šíření takovýchto zpráv je často nelegální.

Ne všechny řetězové emaily musí být nutně škodlivé, může se jednat i například o vtip, obrázky krajin, zvířat anebo různé výzvy k přeposlání ostatním uživatelům, které při splnění slibují odesílateli štěstí. Takovýto typ zpráv ovšem mívá menší dosah než ty, které jsou nějakým způsobem závadné. [7, 21]

Šíření řetězových zpráv není žádnou novinkou, která by přišla s nástupem internetu. První známé instance se objevují již v 19. století v podobě papírových dopisů. Jedním z příkladů je takzvaný “pošli deseticent“ dopis (*send a dime letter*), který vykazoval znaky podobné dnešním emailům. Na začátku byl seznam šesti adres příjemců. V textu byl adresát požádán, aby zaslal deseticent na první adresu v seznamu, poté ji smazal, na konec připojil své kontaktní údaje a přeposlal dopis pěti dalším známým. Na konci mu bylo slíbeno 1 562,5 dolaru. Tato částka vychází z předpokladu, že všichni příjemci instrukce splní a vytvoří se tak strom s 15625 () listy, kde každý list reprezentuje jednoho člověka, který má na vrcholu seznamu původního příjemce, který má dostat oněch 10 centů. Tato struktura mimo jiné odpovídá pyramidovému schématu, které je dnes ilegální. Nejvíce se tento typ dopisu rozšířil v roce 1935 ve Spojených Státech, během ekonomické krize, kde vznikly kopie v řádech desítek milionů. Objevoval se však v různých permutacích po celém světě. [26, 27]

Při zmínce jakékoliv elektronické komunikace se nabízí porovnání se sociálními sítěmi. Propagace informací na nich však probíhá velmi odlišně. Hlavním rozdílem je, že v případě e-mailu se vždy jedná o komunikaci jeden na jednoho. Je samozřejmě možné poslat tu samou zprávu více příjemcům i skupinově odpovídat, ale odesílatel emailu vždy specifikuje příjemce a má plnou kontrolu nad tím ke komu se dostane první instance zprávy a k jejímu dalšímu šíření je potřeba, aby ji příjemce přeposlal dál. Sociální sítě tento typ komunikace umožňují také, ale zároveň jejich uživatelé mohou sdílet příspěvky všem veřejně, případně podmnožině uživatelů v rámci různých zájmových skupin. V tomto případě nemá autor příspěvku 100 % kontrolu nad tím, kdo se k němu dostane. Tento mechanismus pak výrazně zmenšuje vzdálenost putování informací, jelikož samotný veřejný příspěvek funguje jako velký centrální uzel, což u emailu technicky není možné. Propagaci k uživatelům také často obstarávají různé algoritmy a autor nemusí dostat zpětnou vazbu o tom, ke komu se jeho příspěvek dostal.

## Aktuální obsah řetězových emailů v České republice

Aktuálním obsahem řetězových emailů v České republice se zabývá hnutí Čeští elfové. Jedná se o občanské hnutí, které se snaží bojovat proti dezinformačním kampaním na českém internetu a analyzovat je. Členové sami sebe popisují jako patrioty a ty, kteří odmítají přihlížet dezinformačním kampaním cizích států. Věří v české ústavní hodnoty a v členství České republiky v Severoatlantické alianci a Evropské unii. [6]

Toto hnutí pravidelně zveřejňuje analýzy české dezinformační scény, ve kterých se mimo jiné zabývá i obsahem řetězových emailů, zejména těch s politickým obsahem, které sledují zájmy Ruské federace a Čínské lidové republiky.

V prosinci roku 2022 a lednu 2023 se jedním z ústředních témat staly prezidentské volby, pokračujícím tématem byla rusko-ukrajinská válka. Mezi nejvíce napadané subjekty patřili Ukrajina, Petr Fiala a Petr Pavel a česká vláda. Mezi nejvíce podporovanými subjekty byli Rusko, Vladimír Putin a Andrej Babiš. Téma prezidentských voleb v lednu upozadilo déle dominující téma války na Ukrajině. V lednu se také konaly obě kola prezidentských voleb. V únoru 2023 po skončení voleb toto téma vymizelo. Útoky na Petra Pavla a podpora Andreje Babiše mírně oslabily, ale stále se jednalo o jedny z hlavních subjektů zmiňovaných v řetězových emailech. V březnu 2023 se v dezinformačních emailech začal méně objevovat Petr Pavel a cílem se více staly všeobecné sociální a ekonomické problémy obyvatelstva.

Nejčastěji jsou v současnosti řetězové emaily používány k systematické podpoře Ruska s výjimkou podpory Andreje Babiše, kde se čeští elfové domnívají, že jsou zde i jiné vlivy než proruská agenturní síť, nespecifikují však jaké. [8,9,10]

Obsah obrázku text, Písmo, účtenka

Popis byl vytvořen automaticky

*Příklad témat řetězových emailů z března 2023 [10]*

## Vysvětlení odborných pojmů (skok?)

Pro potřeby této práce je nutné definovat několik pojmů, které jsou zásadní pro modely šíření. Jedná se převážně o základy teorie sítí.

### Centrální uzel

Jako centrální uzel, anglicky *hub node*, se označuje takový uzel, který má vysoké množství sousedů*.* Počet sousedů, který definuje toto označení nemusí být pevně definovaný. [4] Pro potřeby implementace algoritmů v praktické části této práce jsem jej nastavil na 50. Centrální uzly jsou zásadní pro generování emailů s velikým dosahem. Pokud se zároveň jedná výchozí bod šíření, dramaticky se pak navyšuje počet příjemců pro všechny modely.

**1.3.2 Sdružovací koeficient – terminologie!**

Sdružovací koeficient uzlu je vlastnost, která popisuje pravděpodobnost propojení jeho sousedů mezi sebou. Je definována jako , kde reprezentuje počet hran mezi sousedy uzlu a je stupeň uzlu.

Sdružovací koeficient sítě *C* je definován jako průměr všech sdružovacích koeficientů uzlů v síti. [22]

Obsah obrázku diagram, Barevnost, design

Popis byl vytvořen automaticky

**kvalita obr**

**1.3.3 Bez-škálová síť**

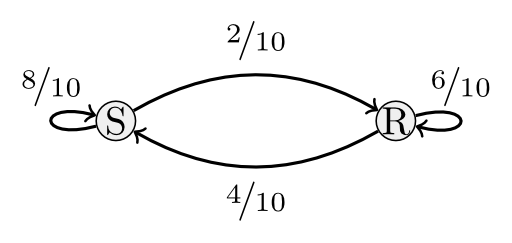
Bez-škálová síť, anglicky *scale-free network*, se vyznačuje vysokým počtem uzlů, které mají malé množství navazujících hran, typicky 1. Zároveň se většina hran koncentruje okolo několika málo centrálních uzlů. Rozložení stupňů uzlů je tedy nerovnoměrné a přibližně se odvíjí od Paretova rozdělení.

Pro tyto sítě platí následující vztah, lépe popsat, stručné, vzorec, distribuce

kde *k*je stupeň uzlu. [4]

**1.3.4 Markovův řetězec**

Markovův řetězec je druh náhodné procházky, kde změna stavu ve stavovém prostoru závisí čistě na aktuálním stavu a pravděpodobnosti nového cílového stavu, nikoliv však na jakémkoliv z předchozích stavů. (12)



*Vizualizace dvoustavového Markovova řetězce [13]*

**1.3.5 Interaktivní Markovův řetězec**

Interaktivní Markovův řetězec je varianta Markovova řetězce, kde kromě jednoduché pravděpodobnosti hraje roli při změně stavu člena systému také stav ostatních členů systému. (14) nechápu

# Modely šíření řetězových emailů

## Všeobecné modely šíření fám

**2.1.1 Maki-Thompson a Daley-Kendall modely první**

Všeobecnými modely, které se nevztahují konkrétně k řetězovým emailům, ale mohou pro ně mít zajímavé implikace se zabývají vědci už delší dobu. Mezi první takové modely patří Daley-Kendall model z roku 1965 a Maki-Thompson model, který jej rozšiřuje, z roku 1973. Tyto modely neberou v potaz topologii reálné sítě anebo používají její velmi zjednodušenou strukturu a dívají se na fámu jako infekci mysli a používají podobné principy jako epidemiologické modely. [11]

Tyto původní modely rozdělují populaci v kontextu jedné fámy na tři skupiny. První skupinou jsou *neznalí*, tedy ti, kteří fámu nikdy neslyšeli. Dále jsou zde *šiřitelé*, ti, kteří se snaží “nakazit“ ostatní a nakonec *potlačovatelé*, ti, kteří už s fámou přišli do styku, dříve ji šířili, ale přestali s tím. Členové skupin mění své zařazení podle určitých pravidel, například pokud přijde neznalý do kontaktu s šiřitelem, okamžitě se nakazí anebo pokud se šiřitel potká s jiným šiřitelem a zjistí, že informaci již zná, tak ztratí zájem o další šíření a stane se z něj potlačovatel. [11]

Modely jsem se rozhodl neimplementovat, jelikož jsou zastaralé, a i když mohou být funkční v malém sociálním kruhu, nemohou plně popsat způsob propagace informací ve velké sociální síti či šíření řetězových emailů. I přes tyto nedostatky se ale staly důležitým stavebním kamenem, na kterém staví modernější studie.

**2.1.2 Model šíření fám v komplexních sociálních sítích**

V roce 2008 se pokusila skupina výzkumníků ve studii *Theory of rumour spreading in complex social networks* popsat všeobecný model šíření fám pomocí rozšíření výše popsaných základních modelů a interaktivních Markovovo řetězců, kde pravděpodobnost změny stavu uzlu závisí také na stavu sousedních uzlů. Typy stavů jsou převzaté z výše zmíněných základních modelů. Zavedli také mechanismus zapomínaní, tedy když se neznalý dostane do styku s informací, ale zapomene ji, dále ji nešíří, přeskočí stav šiřitele a rovnou se z něj stane potlačovatel. [11]

Pomocí aplikací těchto řetězců na menší autoři odvodili sadu rovnic, které popisují průměrné změny stavu každého uzlu pro velké sítě. Tyto rovnice pak aplikovali na různé typy sítí, například náhodnou síť a bez-škálovou síť s exponentem . Jelikož distribuce uzlů náhodné sítě se řídí Poissonovo rozdělením a ne Paretovým, nemá tak velké centrální uzly. [15] Tato vlastnost způsobuje, že náhodné sítě jsou mnohem méně náchylné na šíření nemocí i fám, což zmiňují i sami autoři studie.

Model se zaměřuje spíše na všeobecné šíření informací v síti než konkrétně na řetězové emaily, ale díky tomu, že bere v potaz strukturu sítě, po které se šíří, tak by bylo pomocí něj možné aproximovat i propagaci řetězového emailu. Tento model nabízí opět čistě matematický pohled, který nebere v potaz osobní vztahy mezi odesílatelem a příjemcem. Jeho nevýhodou je také velmi vysoká výpočetní a implementační náročnost. Tento velmi propracovaný model, jsem se rozhodl neimplementovat kvůli jeho přílišné všeobecnosti a velmi vysoké komplexitě. Jedná se však o zásadní výzkum, který ukazuje na možnost existence unifikovaného modelu šíření po síti.

## 2.2 Liben-Nowell a Kleinberg model šíření

David Liben-Nowell a Jon Kleinberg ve své studii *Tracing information flow on a global scale using internet chain-letter* *data* [1] zkoumali šíření řetězových emailů. Jednalo se o několik internetových petic, konkrétně petice proti válce v Iráku z let 2002 a 2003 a pro zachování vládní podpory pro veřejnoprávní rádio NPR ve Spojených Státech z roku 1995. Data byla sesbírána z veřejného internetového archivu emailů.

Za normálních okolností je téměř nemožné legálně vystopovat cestu řetězového emailu. Jelikož na rozdíl od sociálních sítí emailové zprávy jsou soukromé a příjemce emailu se musí sám dobrovolně rozhodnout, zda zprávu zveřejní. Petice je však unikátní v tom, že stačí relativně malé procento příjemců, které zprávu zveřejní pro rekonstrukci šíření. Toto je dáno tím, že lidé, kteří zprávu nezveřejní v celé své podobě, ale chtějí petici podpořit, připojí své jméno jako podpis k emailu a přepošlou jej dál. Jejich jméno pak může být zveřejněno někým jiným.

### Rekonstrukce šíření petice z reálných dat [28]

### Autoři v první fázi výzkumu hledali kopie petic pomocí různých internetových vyhledávačů. Jako vyhledávací parametr použili jméno a bydliště prvního signatáře. Takto získali 1265 unikátních webových stránek. Nalezené webové stránky autoři prozkoumali a vyřadili všechny stránky, kde se nejednalo o kopii petice. Mezi vyřazenými stránkami byly například různé pojednání o petici anebo rozhovory s prvním signatářem petice. Po vyřazování zbylo 1132 kopií petice, které však obsahovaly duplicity, jelikož některé unikátní kopie byly zveřejněny na více webových adresách.

### Z těchto sesbíraných kopií rekonstruovali autoři síť šíření petice pomocí následujících kroků:

### Pro každý seznam signatářů a každou po sobě jdoucí dvojici podpisů typograficky standardizovali text. Všechny neabecední znaky, kromě mezer byly odstraněny a všechna jména příjemců převedli na malá písmena. Následně zaznamenali spojení ze jména do jména .

### Pro každé unikátní jméno prvního signatáře ve všech kopií petice vytvořili číselný identifikátor a každé jméno, které se shodovalo s tímto unikátním jménem přiřadili k jeho identifikátoru. Po zpětné ruční kontrole autoři zjistili, že všechny kopie petice měly stejného prvního signatáře, tedy i stejný identifikátor .

### Jako je označena množina všech jmen , pro které je zaznamenáno spojení ke jménu přiřazenému k identifikátoru . Jako editační vzdálenost je označen minimální počet vložení, nahrazení anebo smazání jednoho znaku, které jsou nutné pro transformaci textového řetězce na textový řetězec .

### Pro každé je zaznamenána hrana v grafu a identifikátor , který je určen pomocí následujících mechanismů:

### Pokud je jméno y identické se jménem, ke kterému již byl přiřazen jiný identifikátor , potom

### Pokud je jméno v editační vzdálenosti 6 od jména , kterému již byl přiřazen identifikátor , potom

### Pokud neplatí žádná z výše uvedených podmínek, pak je jménu přiřazen nový identifikátor

Struktury vytvořené z dat pro zkoumané petice pomocí výše uvedeného postupu jsou hlubokým stromem s nízkým počtem rozvětvení, což je v rozporu s výsledky ostatních modelů, založených na epidemiologickém šíření [1,11]

Tento přístup má tu výhodu, že se zakládá na 100 % reálných datech. Nevýhodou je, že tato data nejsou úplná, z principu nelze zachytit všechny instance přeposlání zprávy a také nepopisují přesně strukturu emailové sítě.

### Model šíření petic (LNK model)

Tento model se odlišuje od ostatních modelů [11] tím, že je odvozen od šíření konkrétních emailů. Většina ostatních modelů také nepředpokládá, že emailová síť je něco specifického a dívá se na ni pouze jako na další sociální anebo všeobecnou síť. Vzhledem k tomu, že reálné sítě, včetně sociálních, jsou většinou bez-škálové [15,20], šíření v nich probíhá expanzivně a v malém počtu kroků, ale s velkým dosahem. Toto je způsobeno existencí centrálních uzlů s extrémně vysokým počtem sousedů, které zkracují vzdálenosti mezi jednotlivými uzly. Velmi důležitým faktem, který zdůrazňují i sami autoři studie je, že výsledná struktura propagace petice v LNK modelu velmi připomíná hluboký strom s malým počtem rozvětvení. vysvětlit Tato struktura neodpovídá ostatním modelům, kde průměrné vzdálenosti mezi uzly jsou malé.

Mají více modelů - rozdělit Model popisuje reakci příjemců na zaslaný email. Příjemce s určitou pravděpodobností na email reaguje anebo jej ignoruje. Toto je reprezentováno pravděpodobností ignorování emailu , která se pohybuje okolo 0.65, ale její přesná hodnota v rozumném rozsahu od 0.5 do 0.75 nemá zásadní vliv na výslednou strukturu.

Pokud se příjemce rozhodne reagovat, tak má dvě vzájemně se vylučující možnosti reakce. První a častější variantou, je že na něj na pouze skupinově odpoví odesílateli a všem spolupříjemcům a připojí své jméno k petici s pravděpodobností , ale nepřepošle ji dále. Hodnota se v simulacích autorů typicky pohybovala okolo 0.95. Druhá možnost, která nastává s pravděpodobností , že příjemce rozešle kopii s připojeným podpisem všem svým sousedům. Vysoká pravděpodobnost skupinové odpovědi napomáhá tomu, že výsledná struktura ve výsledku připomíná hluboký strom obrázek a přibližně odpovídá rekonstruovaným datům. Čím vyšší je hodnota tohoto parametru, tím je menší šance na rozšíření petice k novým adresátům, jelikož skupinová odpověď pouze napomáhá připojení odpovídajícího jednotlivce do zveřejněného seznamu podpisů.

Pokud je petice přeposlána dál, je pak společně se všemi připojenými podpisy zpřístupněna ve veřejném archivu emailů s pravděpodobností , jejíž hodnota se pohybuje okolo 0.22. Tento parametr způsobuje, že pokud se má ve výsledcích zobrazit konkrétní příjemce, musí on anebo někdo, komu petici přeposlal se svým připojeným jménem, zveřejnit. Tento parametr je zásadní, jelikož simuluje pozorovatelnost reálných výsledků. V realitě nikdy není možné získat všechny kopie řetězového emailu s větším dosahem a vždy se pracuje pouze s nějakou zveřejněnou podmnožinou. Výhodou petice je motivace podepsaných zvyšovat povědomí o jejím obsahu, což z ní děla vhodný objekt ke zkoumání.

Unikátní vlastností tohoto modelu je také fakt, že reakce na zprávu probíhá v reálném čase, ne v okamžiku přijetí. Při přijetí zprávy se v případě, že se příjemce rozhodnul reagovat, vygeneruje podle distribuční funkce čas *t*. Distribuční funkce vychází ze studií o rychlosti lidských odpovědí v komunikaci a její konkrétní hodnota použitá pro LNK model je , kde . Příjemce poté čeká po dobu *t,* než odpoví. Během této doby mu může přijít více stejných emailů, reaguje však pouze na jeden, ten, který v sobě obsahuje největší množství jmen. V reálném světě komunikace také neprobíhá okamžitě, ale trvá určitý čas, než příjemce zareaguje.

Během svého výzkumu autoři studie implementovali variantu modelu, který nezakomponoval reakce v reálném čase. Výsledky simulací z této varianty však neodpovídaly struktuře získané z reálných petic a podobaly se spíše ostatním modelům, tedy široce rozvětvenému grafu, kde vzdálenosti mezi uzly byly relativně malé. Implementace varianty s reálnými časy odpovědí byla jedním, ze dvou kroků, která byla nutná k dosažení kýžené stromové struktury pro sítě s vysokým sdružovacím koeficientem. Vysoký sdružovací koeficient mají mimo jiné i bez-škálové sítě. Druhým krokem bylo zavedení parametru odpovědi .

Reálná data ukazují, že většina příjemců má tendenci přeposlat email pouze jednou, či několika známým [1,21]. Tento fakt je v LNK modelu reprezentován malou šancí na přeposlání emailu dál.

Existují také tzv. super-šiřitelé, kteří jsou schopni zaslat email na řádově stovky adres. Tento fakt přímo v LNK modelu zohledněn není, ale je reprezentován jako centrální uzel v síti. [21]

Autoři pro simulaci nepoužili emailovou síť, ale reprezentaci několika velkých sociálních sítí, z nichž největší byla sociální síť LiveJournal. info

Vyvstává tedy otázka, proč jsou výsledky modelu z této studie tolik odlišné od ostatních modelů. Nejpravděpodobnějším vysvětlením je, že model je specifický pro všeobecné šíření petic, či dokonce pouze pro šíření konkrétních petic, které autoři zkoumali, ne však všech řetězových emailů. Sami autoři uvádí, že stromová struktura je způsobená realistickým časem odpovědi a parametrem . Čas odpovědi je naprosto relevantní pro všeobecné řetězové emaily. Tomu tak pravděpodobně není u parametru, jeho základní mechanismus lze sice aplikovat, ale u běžného emailu není typicky k dispozici historie podpisů, tedy lidí, ke kterým se zpráva dostala v minulosti a při aplikaci na běžný email, nastane efekt, že ten, kdo email nepřepošle k novým příjemcům se nedostane do viditelné struktury, ale pouze vytváří jeho duplicitu v schránkách spolupříjemců, jelikož pokud po něm někdo zveřejní kopii emailu, neexistuje v emailu záznam o tom, že se k němu dostal i ten co pouze odpověděl a nešířil jej k novým příjemcům.

Struktura šíření ostatních modelů je často odvozená od šíření informací po sociálních sítích anebo dokonce od epidemiologických modelů. Sociální sítě jsou mnohem lépe dokumentovatelné díky jejich veřejné povaze. Žádná struktura šíření ostatních modelů ani vzdáleně nepřipomíná strom. Rozdílné jsou však i struktury sítí. Sociální sítě jsou typicky bez-škálové, ale u emailové sítě je toto tvrzení komplikovanější. proč?

## 2.3 Galton – Watson model

Pro LNK model jsou typické hluboké stromy s nízkou rozvětveností, které tento algoritmus generuje. Tuto strukturu lze replikovat Galton – Watsonovým modelem. [4] Jedná se o překvapivě velice jednoduchý postup pro generování stromů, který byl publikován již v 19. století a jeho původní účel nemá nic společného s emaily ani šířením zpráv. Sloužil k simulaci historie příjmení v populaci.

Algoritmus generuje stromy pomocí pravděpodobnostního rozdělení počtu potomků. Založí graf s jedním uzlem, který označí jako kořen stromu a náhodně vygeneruje počet potomků, podle předem dané distribuce. Mezi rodiče a potomky vloží hranu a proces opakuje pro každý další vložený uzel. Proces končí po dosažení předem daného počtu generací anebo pokud všechny uzly aktuální generace mají počet potomků roven 0. [5]

Jak funguje, kde se tabulka vzala? Pro původní LNK model, kde *k* se rovná počtu dětí je distribuce následovná. [25]

Obsah obrázku text, snímek obrazovky, řada/pruh, Písmo

Popis byl vytvořen automaticky

Opakované běhy s výše uvedenou distribucí, také dobře vizualizují skutečnost, že rozšíření e-mailu velkému množství příjemců není nemožné, ale je velmi málo pravděpodobné. Většina běhů skončí s malým počtem uzlů, před dosažením limitu počtu generací, pokud je tento limit přibližně nastaven alespoň okolo hodnoty 100.

Výše uvedená distribuce však nemusí plně zachycovat fenomén super-šiřitelů, kteří mívají řádově až stovky odeslaných emailů a jsou často původními tvůrci obsahu. I když je potřeba mít na paměti, že LNK model, se zabývá pouze viditelnou podmnožinou šiřitelů, je velmi nepravděpodobné, že by tato podmnožina byla zredukována na maximálně 3 potomky startovního uzlu, který je v tomto případě zároveň uzlem centrálním.

Toto nastavení by mohlo indikovat, že obě zkoumané petice v LNK modelu jsou vzácným případem řetězového emailu s velkým dosahem, který se rozšířil spontánně, bez umělého startovního impulsu super-šiřitele.

Tento model jsem se rozhodl implementovat, i když jej asi nelze považovat za plnohodnotnou simulaci šíření řetězových emailů. Nabízí však zajímavou perspektivu a je schopen věrohodně napodobit výsledky původního LNK modelu. Jeho výhodou je také nízká výpočetní náročnost, díky které je možné získat velké množství referenčních dat.

# Implementace simulace emailové sítě a modelů šíření

## Použitá testovací data a způsob jejich zpracování

Stručné!Pro účely této práce jsem použil existujících data emailových sítí, data o uživatelích slovenské sociální sítě Pokec a vygenerovanou bez-škálovou síť pomocí Barabási-Albertova Alberts? algoritmu. [16]

U emailových sítí jsem se zaměřil na data z nespecifikované velké výzkumné evropské organizace. Anonymizovaná data pochází z října 2003 až května 2005. Z přibližně 3 milionů emailů byl sestaven graf ve kterém každý uzel reprezentuje unikátní emailovou adresu a každá hrana mezi adresami reprezentuje alespoň jeden vyměněný email. [2,3]

Použil jsem jak kompletní síť, tak její podmnožinu [17]. Kompletní síť má výhodu větší přesnosti výsledků, podmnožina umožňuje provést více simulací šíření ve stejném čase.

Všeobecně je získání reálných dat emailových sítí velmi složité, díky nutnosti souhlasu všech účastníků komunikace. Z tohoto důvodu jsem zvolil takto relativně starý data set, jelikož jako jeden z mála veřejných data setů odpovídal potřebám této práce.

Data profilů sociální sítě Pokec [18,19] jsem zvolil kvůli tomu, že obsahuje rozsáhlá data uživatelských profilů. Sice se nejedná o emailovou síť, ale pokud jsem chtěl do práce zakomponovat i model založený na demografických datech, neměl jsem jinou možnost, jelikož žádná veřejná emailová síť s takto rozsáhlými daty neexistuje.

Jména uživatelů byla anonymizována, ale všechna ostatní data, která uživatelé dobrovolně zveřejnili, včetně věku, byla zachována. Uzel reprezentuje uživatelský profil, hrana pak reprezentuje vztah “přátelství” mezi profily, podobný vztah mezi uživateli používá například síť Facebook.

Pro potřebu Relatability modelu založeném na věku, bylo potřeba doplnit věk pro všechny uživatelské profily, jelikož ne všichni tuto informaci zveřejnili. Věk jsem dopočítával jako průměr věku všech přátel, kteří tuto informaci zveřejnili, jelikož lze předpokládat, že lidé preferují komunikaci ve své věkové skupině.

Kvůli obrovské velikosti sítě jsem vytvořil také několik podmnožin. Vytvořil jsem algoritmus, který ze sítě náhodně vybral jeden uzel a postupně prošel všechny jeho sousedy a přidal je do nového grafu, toto se opakovalo pro všechny nově přidané uzly, dokud počet uzlů v grafu nepřekročil předem danou vstupní hodnotu.

Barabási-Albert síť jsem vygeneroval pomocí knihovní funkce Python modulu networkx.

Všechny reálné sítě byly reprezentovány jako seznam hran, kde každá hrana byla reprezentována jako dva uzly, mezi kterými se nachází, tyto seznamy jsem načetl, zrekonstruoval z nich síť a vyexportoval je do formátu gexf. Takto vyexportované grafy jsem načetl do aplikace Gephi. Pomocí této aplikace jsem nastavil pozici x a y pro každý uzel, pro potřeby vizualizace. Následně jsem data vyexportoval v json formátu a nahrál je do své aplikace. Tato vyexportovaná data jsou součástí přílohy této práce.

## 

## Návrh a implementace modifikovaného LNK modelu

Cílem práce je zkoumání všeobecných řetězových emailů a u LNK modelu předpokládám, že je příliš specifický pouze pro šíření petic, kvůli parametru skupinové odpovědi. Model je však přesto unikátní a pro účely této práce zajímavý, a proto jsem se rozhodl navrhnout a implementovat jeho upravenou variantu. Pro porovnání výsledků s originálním LNK modelem jsem použil referenční Galton-Watsonův model generování stromů. [4] Co se vlastně implementuje?

Pro výběr počátečního uzlu jsem naimplementoval dva mechanismy, prvním je výběr náhodného uzlu v grafu, čímž se snažím reprezentovat spontánní vzniknutí zprávy, kde kdokoliv z jakéhokoliv důvodu začne šířit obsah, který mu přijde zajímavý. Druhým mechanismem je start z náhodného centrálního uzlu s alespoň 50 sousedy, který simuluje záměrné šíření s cílem dostat zprávu k co největšímu počtu lidí. Tento způsob je typický pro šíření nepravdivé propagandy cizích států.

Parametr , pravděpodobnost ignorování zprávy, jsem ponechal beze změny a pro simulace jsem jeho hodnotu ponechal ve stejném rozsahu jako v původním modelu, tedy 0.5 až 0.75. Toto vychází z domněnky, že v tomto ohledu by neměl být zásadní rozdíl mezi peticí a všeobecným řetězovým emailem.

Beze zásadní změny jsem ponechal i , pravděpodobnost zveřejnění emailu. Jednou ze zajímavých vlastností tohoto modelu, je fakt, že nesimuluje absolutní dosah zprávy, ale pouze její pozorovatelnou část. Hodnoty jsem ponechal v podobném rozsahu, tedy 0.20 až 0.25. Tento rozsah se může zdát na první pohled nepřirozeně vysoký. U petice se dá předpokládat, že se ji lidé primárně snaží šířit z dobré vůle, jelikož upřímně věří jejímu obsahu, a proto ji zveřejňují. U motivace zveřejnění řetězového emailu záleží na jeho obsahu, ale trendy poslední doby, alespoň v České republice ukazují, že obsah řetězových emailů často útočí na emoce a šíří nepravdivou politickou propagandu cizích států. [7,8,9,10] Domnívám se, že toto má za následek dva druhy motivace pro zveřejnění. První je typický pro lidi s proti-systémovým smýšlením, kteří mají tendenci obsahu emailu věřit a snaží se o něm zvýšit povědomí, jelikož obsažené informace vnímají jako pravdu utajovanou státními institucemi a mainstreamovými médii. Druhý druh motivace je naopak typický pro lidi, kteří se proti těmto emailům snaží bojovat a vyvracet je. Zveřejněním, například v databází řetězových emailů Eldariel [23], zvyšují osvětu o těchto hoaxech. Parametr reprezentuje motivaci pro sdílení, jeho hodnoty i jeho použití jsou tedy stejné i v mé implementaci LNK modelu, ale důvodně se domnívám, že důvody pro zveřejnění se mohou částečně lišit.

Simulace odpovědi v reálném čase a generování parametru *t* probíhá velmi podobně jako v originálním modelu, podle stejné distribuční funkce. Jediným rozdílem je, že čas neplyne kontinuálně, ale ve velmi malých diskrétních krocích, které toto plynutí simulují. Tento implementační přístup, při zvolení dostatečně malých kroků, zásadně neovlivňuje celkové výsledky modelu. Tyto malé kroky minimalizují pravděpodobnost aktivace dvou uzlů ve stejný čas, což je důležité pro vytváření dlouhých emailových řetězců.

Nejvýraznějším rozdílem mezi originálním modelem a mojí implementací je pojetí parametru. U obecného řetězového emailu není dostupný seznam podpisů, takže i když tento parametr reprezentuje to samé, jeho implikace je velmi rozdílná. Pokud se příjemce rozhodne reagovat s pravděpodobností , pak se s pravděpodobností rozhodne pouze skupinově odpovědět. Hodnoty této pravděpodobnosti jsem použil podobné jako autoři LNK modelu, konkrétně v intervalu od 0.88 do 0.95. Skupinová odpověď ovšem nemá v tomto případě efekt vytváření dlouhých řetězců propagace, protože díky absenci podpisů, nemůže jméno odpovídajícího zveřejnit někdo jiný a nikdy se nedostane do pozorovatelné struktury šíření. Z tohoto faktu vyplývá, že reálná šance na přeposlání je rovna . Což pro parametry a vychází na přibližně 0.03, tedy 3%. Tato hodnota je ve shodě s tvrzením Josefa Šlerky, předního českého odborníka na online komunikaci z Univerzity Karlovy, který v rozhovoru pro Český rozhlas uvedl, že řetězové emaily v České republice přeposílají zhruba 3% příjemců [24].

Po dokončení simulace se z výsledků rekonstruuje podmnožina šiřitelů. Pro každý uzel, který zprávu zveřejnil s pravděpodobností , je zrekonstruována cesta ke startovnímu uzlu pomocí Dijkstrova algoritmu. Proč? Díky tomuto mechanismu je možné pozorovat ucelenou strukturu šíření, jelikož takto se do výsledného grafu dostanou i některé uzly, které zprávu přeposlaly, ale nezveřejnily. Pokud bych tento mechanismus odebral, výsledkem by nebyl spojitý graf, ale pouze nespojité ostrůvky, řádově v jednotkách uzlů. Taková struktura by sice více odpovídala reálným pozorovaným datům, ale neměla by vypovídající hodnotu pro pozorování struktury šíření.

Tento model má za cíl simulovat strukturu šíření emailu, nikoliv jeho absolutní dosah. Jeho výhodou je, že počítá s nekompletností pozorovaných dat a je částečně odvozen od reálných pozorování. Nevýhodou je, že nepočítá s demografickými ukazateli ani s osobními vztahy příjemce a odesílatele.

# Závěr

Tato kapitola se nečísluje. Rozsah je zpravidla 5-10 normostran.

# Seznam použitých zdrojů

1. Liben-Nowell, D., & Kleinberg, J. (2008). Tracing information flow on a global scale using Internet chain-letter data. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(12), 4633–4638. Dostupné z: <https://doi.org/10.1073/pnas.0708471105>
2. *Stanford Network Analysis Project: SNAP* [online]. [cit. 2023-06-03]. Dostupné z: <http://snap.stanford.edu/data/email-EuAll.html>
3. J. Leskovec, J. Kleinberg and C. Faloutsos. [Graph Evolution: Densification and Shrinking Diameters](http://www.cs.cmu.edu/~jure/pubs/powergrowth-tkdd.pdf). ACM Transactions on Knowledge Discovery from Data (ACM TKDD), 1(1), 2007.
4. Golub, B., & Jackson, M. O. (2010). Using selection bias to explain the observed structure of internet diffusions. *Proceedings of the National Academy of Sciences*, 1–1. <https://doi.org/10.1073/pnas.1000814107>
5. Branching processes, University of Chicago. *Stat.uchicago.edu* [online]. [cit. 2023-06-10]. Dostupné z: <http://galton.uchicago.edu/~lalley/Courses/312/Branching.pdf>
6. Čeští elfové. *Cesti-elfove.cz* [online]. [cit. 2023-06-02]. Dostupné z: <https://cesti-elfove.cz/uvodni-strana/>
7. Čeští elfové. *Cesti-elfove.cz* [online]. [cit. 2023-06-02]. Dostupné z: <https://cesti-elfove.cz/retezove-e-maily/>
8. Čeští elfové, měsíční report prosinec 2022, leden 2023. *Cesti-elfove.cz* [online]. 23.2.2023 [cit. 2023-06-02]. Dostupné z: <https://cesti-elfove.cz/wp-content/uploads/MM_2022_01.pdf>
9. Čeští elfové, měsíční report únor 2023. *Cesti-elfove.cz* [online]. 13.4.2023 [cit. 2023-06-02]. Dostupné z: <https://cesti-elfove.cz/wp-content/uploads/MM_2023_02.pdf>
10. Čeští elfové, měsíční report březen 2023. *Cesti-elfove.cz* [online]. 13.4.2023 [cit. 2023-06-02]. Dostupné z: <https://cesti-elfove.cz/wp-content/uploads/MM_2023_03.pdf>
11. NEKOVEE, M., Y. MORENO, G. BIANCONI a M. MARSILI. Theory of rumour spreading in complex social networks. *Physica A: Statistical Mechanics and its Applications* [online]. 2007 [cit. 2023-06-02]. ISSN 03784371. Dostupné z: <https://doi.org/10.1016/j.physa.2006.07.017>
12. Behrends, E. (2000). Markov chains: how to start? In *Introduction to markov chains: With special emphasis on rapid mixing*. essay, Vieweg.
13. THOMA, Martin. Markov chain. In: *Martin-thoma.com* [online]. 12.5.2015 [cit. 2023-06-02]. Dostupné z: <https://martin-thoma.com/images/2015/05/markov-chain-rain-sun.png>
14. Arnold, F., Gebler, D., Guck, D., & Hatefi, H. (2014). A tutorial on interactive markov chains. *Stochastic Model Checking. Rigorous Dependability Analysis Using Model Checking Techniques for Stochastic Systems*, 4–6. Dostupné z: <https://doi.org/10.1007/978-3-662-45489-3_2>
15. Barabási, A.-L., & Pósfai, M. (2017). Chapter 3 Random Networks. In *Network science*. essay, Cambridge University Press.
16. Barabási, A.-L., & Pósfai, M. (2017). Chapter 5 The Barabási-Albert Model. In *Network science*. essay, Cambridge University Press.
17. Ryan A. Rossi, & Nesreen K. Ahmed (2015). The Network Data Repository with Interactive Graph Analytics and Visualization. In AAAI.
18. L. Takac, M. Zabovsky. [Data Analysis in Public Social Networks](http://snap.stanford.edu/data/soc-pokec.pdf), International Scientific Conference & International Workshop Present Day Trends of Innovations, May 2012 Lomza, Poland.
19. *Stanford Network Analysis Project: SNAP* [online]. [cit. 2023-06-03]. Dostupné z: <http://snap.stanford.edu/data/soc-Pokec.html>
20. Barabási, A.-L., & Pósfai, M. (2017). Chapter 4 The Scale-Free property. In *Network science*. essay, Cambridge University Press.
21. Síť řetězových e-mailů funguje spontánně, dezinformátoři ji ale umí inspirovat a zneužívat, říká Kučík z Českých elfů. *Forum24* [online]. [cit. 2023-06-04]. Dostupné z: <https://www.forum24.cz/vit-kucik-sit-retezovych-mailu-funguje-spontanne-dezinformatori-ji-ale-umi-inspirovat-a-zneuzivat/>
22. Barabási, A.-L., & Pósfai, M. (2017). Chapter 2 Graph Theory. In *Network science*. essay, Cambridge University Press.
23. Čeští elfové, databáze emailů Eldariel. *Cesti-elfove.cz* [online]. [cit. 2023-06-02]. Dostupné z: <https://eldariel.cesti-elfove.cz/>
24. ŠLERKA, Josef. *Jak to vidí….* Rozhlas, Český rozhlas Dvojka, 24.8.2022. Dostupné také z: <https://dvojka.rozhlas.cz/medialni-analytik-josef-slerka-ruska-propaganda-je-strikacka-lzi-ukrajinci-maji-8812711>
25. Golub, B., & Jackson, M. O. (2010). Using selection bias to explain the observed structure of internet diffusions. *Proceedings of the National Academy of Sciences*, 2–2. <https://doi.org/10.1073/pnas.1000814107>
26. SOLLY, Meilan. *Before Chain Letters Swept the Internet, They Raised Funds for Orphans and Sent Messages From God* [online]. 2.6.2020 [cit. 2023-06-10]. Dostupné z: <https://www.smithsonianmag.com/history/chain-letters-swept-internet-they-raised-funds-orphans-and-conveyed-messages-god-180975005/>
27. *The “Send a Dime” Chain Letter craze of the 1930’s* [online]. [cit. 2023-06-10]. Dostupné z: <https://www.geekslop.com/life/fads-and-trends/fads-and-trends-1900/2010/send-a-dime-chain-letter-1930s>
28. Liben-Nowell, D., & Kleinberg, J. (2008). Tracing information flow on a global scale using Internet chain-letter data. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(12), supporting information (SI) Appendix. Dostupné z: <https://doi.org/10.1073/pnas.0708471105>