**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

# 1 Řetězové emaily

## 1.1 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 vrchol (viz kapitola 2.1.1), 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.

## 1.2 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

*Obrázek 1: Příklad témat řetězových emailů z března 2023 [10]*

# 2 Úvod do teorie sítí

Teorie sítí je vědní disciplína vycházející z teorie grafů, která se zabývá popisem reálných systémů skládajících se z mnoha navzájem interagujících komponent. Jedná se velmi novou vědní disciplínu, její rozvoj jako samostatného oboru začal až na začátku 21. století. Na první pohled se může zdát, že název teorie sítí je synonymem pro teorii grafů, která existuje již několik století. Hlavním rozdílem je, že tento vědní obor se nezabývá pouze abstraktními matematickými modely, ale studuje reálné systémy reprezentovatelné pomocí grafu a snaží se pochopit jejich chování. Díky tomuto faktu nabízí teorie sítě vhodné nástroje i pro analýzu síření řetězových emailů. [3]

## 2.1 Vysvětlení odborných pojmů

Pro potřeby této práce je nutné definovat několik pojmů, které jsou důležité pro modely šíření. Jedná se o základní pojmy z teorií sítí a grafů.

### Centrální vrchol

Jako centrální vrchol, anglicky *hub node*, se označuje takový vrchol, který má vysoké množství sousedů*.* Počet sousedů, který definuje toto označení nemusí být pevně daný. [20] Pro potřeby implementace algoritmů v praktické části této práce jsem při použití centrálního vrcholu jako původního vrcholu, ze kterého se šíří zpráva, použil vrchol s největším stupněm v rámci dané sítě. Centrální vrcholy 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. **Shlukovací koeficient**

Pro vrchol je lokální shlukovací koeficient vrcholu vlastnost, která popisuje pravděpodobnost propojení jeho sousedů mezi sebou. Je definována jako , kde reprezentuje počet hran mezi sousedy vrcholu a je stupeň vrcholu. Shlukovací koeficient sítě *C* je definován jako průměr všech shlukovacích koeficientů vrcholů v síti. V angličtině je tento termín pojmenován jako *clustering coeficient.* [22]

Obsah obrázku diagram, Barevnost, design

Popis byl vytvořen automaticky

*Obrázek 2: Ilustrativní obrázek shlukovacího koeficientu*

* 1. **Typy sítí**

**2.2.1 Náhodná síť**

Náhodná síť je typ sítě, ve které je vrcholů a náhodně vytvořených hran, které tyto vrcholy spojuje.

Takovouto síť je možné zrekonstruovat pomocí Erdős-Rényi algoritmu. Algoritmus má dva vstupní parametry. Prvním je , počet vrcholů výsledné sítě. Druhým je , pravděpodobnost vytvoření hrany každou individuální dvojicí vrcholů v síti. Algoritmus má následující kroky:

1. Vytvoří se síť s  vrcholy a 0 hranami.
2. Pro každou z  dvojic vrcholů:
   * Vytvoří se hrana mezi oběma vrcholy z dvojice s pravděpodobností

Pro takto vygenerovanou se rozdělení stupňů vrcholů typicky řídí Poissonovo rozdělením. Lokální shlukovací koeficient jednotlivých vrcholů v těchto sítích nijak nekoreluje s jejich stupněm, což vyplývá z náhodného vkládání hran mezi vrcholy. Tento model byl používán v druhé polovině dvacátého století pro modelování velkých komplexních systémů. S nástupem teorie sítí a pozorováním reálných dat však výzkumníci zjistili, že distribuce vrcholů v reálných sítí se typicky neřídí Poissonovo rozdělením a lokální shlukovací koeficient vrcholu se odvíjí od jeho stupně, čím vyšší stupeň vrcholu, tím bývá vyšší i jeho lokální shlukovací koeficient. Dalším rozdílem oproti reálným sítím je stupeň centrálních vrcholů. Centrální vrcholy v náhodných sítích mívají typicky menší stupeň než centrální vrcholy v reálné síti se stejným počtem vrcholů. Kvůli těmto nesrovnalostem byl navržen model bez-škálových sítí, které lépe popisují chování velkých systémů v reálném světě. [15]

**2.2.2 Bez-škálová síť**

Převážnou část sítí popisující velké reálné systémy lze klasifikovat jako sítě bez-škálové, anglicky *scale-free network*. Tyto sítě se vyznačují vysokým počtem vrcholů, které mají malé množství navazujících hran, typicky 1. Zároveň v sítí existuje několik málo centrálních vrcholů, které mají velmi vysoký stupeň. Rozložení stupňů vrcholů je tedy nerovnoměrné a přibližně se odvíjí od mocninného zákona, anglicky *power law*.

Pro bez-škálové sítě platí následující vztah, *k*je stupeň vrcholu a je pravděpodobnost s jakou bude mít vrchol stupeň alespoň , , kde platí že. Parametr se nazývá stupňovým exponentem. Tento vztah způsobuje výše zmíněný fakt, že většina vrcholů má malý počet sousedů, a naopak několik vrcholů okolo sebe koncentruje vysoký počet spojení.

Stupňový exponent musí být v rozmezí od 2 do 3, aby mohla být velká síť klasifikována jako bez-škálová.

Pro exponenty menší než 2 začne křivka distribuce klesat příliš prudce a pravděpodobnost výskytu centrálních vrcholů téměř neexistuje a zároveň dochází počet vrcholů ke kterému se mohou nové vrcholy v síti připojit, což de facto znemožňuje existenci velké bez-škálové sítě se stupňovým exponentem menším než 2. Pro exponenty větší než 3 klesá křivka distribuce naopak moc pomalu a centrální vrcholy jsou díky tomu menší, ale více početné, díky čemuž je síť v takovémto režimu často nerozeznatelná od sítě náhodné. [20]

Emailové sítě, ve kterých vrcholy reprezentují jednotlivé emailové adresy a hrany reprezentují vyměněné zprávy byly pozorovány jako bez-škálové. [13]

Obrázky 3, 4 a 5 poukazují na různé vlastnosti náhodných a bez-škálových sítí, zelená barva reprezentuje náhodné sítě, fialová barva reprezentuje sítě bez-škálové.

Obsah obrázku text, diagram, řada/pruh, Vykreslený graf

Popis byl vytvořen automaticky

*Obrázek 3: vizualizace distribuce uzlů v náhodných a bez-škálových sítí [20]*

Obsah obrázku kruh, umění

Popis byl vytvořen automatickyObsah obrázku origami

Popis byl vytvořen automaticky se střední mírou spolehlivosti*Obrázek 4: ilustrativní vizualizace náhodné sítě [20*Obsah obrázku kruh, umění

Popis byl vytvořen automatickyObsah obrázku origami

Popis byl vytvořen automaticky se střední mírou spolehlivosti*] Obrázek 5: ilustrativní vizualizace bez-škálové sítě [20]*

Z obrázků je patrné, jak je v náhodné síti rozložení hran relativně rovnoměrné, a naopak v síti bez-škálové se propojení vrcholů koncentruje okolo několika centrálních vrcholů.

# Modely šíření řetězových emailů

## 3.1 Všeobecné modely šíření fám

**3.1.1 Maki-Thompson a Daley-Kendall modely [11]**

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, dívají se na fámu jako “infekci mysli“ a používají podobné principy jako epidemiologické modely. Tyto modely lze aplikovat na všeobecné sociální kruhy, nemusí se nutně jednat o velkou emailovou síť.

Tyto původní modely rozdělují populaci v kontextu jedné fámy na tři skupiny. První skupinou jsou *neznalí (ignorants),* tedy ti, kteří fámu nikdy neslyšeli. Dále jsou zde *šiřitelé (spreaders)*, ti, kteří se snaží “nakazit“ ostatní a nakonec *potlačovatelé (stiflers)*, ti, kteří už s fámou přišli do styku, dříve ji šířili, ale přestali s tím.

Kontakt v těchto modelech probíhá vždy po párech, kde jedním z páru je vždy *šiřitel*. V Daley-Kendall modelu se nerozlišuje mezi odesílatelem a příjemcem, v Maki-Thompson modelu jsou odesílatel a příjemce od sebe rozlišeni.

V Daley – Kendall modelu mění členové skupin své zařazení podle následujících pravidel:

* Pokud přijde *neznalý* do kontaktu s *šiřitelem*, *neznalý* se okamžitě “nakazí“ a stane se z něj *šiřitel* s pravděpodobností .
* Pokud se *šiřitel* potká s jiným *šiřitelem* anebo *potlačovatelem,* oba zjistí, že informace je již známá, ztratí zájem o její další šíření a z obou členů párů, kteří jsou *šiřiteli* se stane *potlačovatel* spravděpodobností.

V Maki – Thompson modelu mění členové skupin své zařazení podle následujících pravidel:

* Pokud přijde *neznalý* do kontaktu s *šiřitelem*, *neznalý* se okamžitě “nakazí“ a stane se z něj *šiřitel* s pravděpodobností . Tento krok je identický i v Daley – Kendall modelu
* Pokud se *šiřitel* potká s jiným *šiřitelem* anebo *potlačovatelem,* zjistí, že informace je již známá, ztratí zájem o její další šíření a z iniciujícího *šiřitele* se stane *potlačovatel* spravděpodobností. Tento krok je rozdílný od druhého kroku v Daley – Kendall modelu, protože od sebe rozlišuje odesílatele a příjemce fámy.

Tyto dva modely jsem se rozhodl neimplementovat, jelikož jsou příliš jednoduché, 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.

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

V roce 2008 se pokusila skupina výzkumníků ve studii *Theory of rumour spreading in complex social networks* [11] popsat všeobecný model šíření fám pomocí rozšíření Maki – Thompson modelu. Autoři zde používají stejné rozdělení populace do skupin jako Daley-Kendall a Maki-Thompson modely, tedy *neznalé, šiřitele a potlačovatele.* Zavádí však mechanismus spontánního zastavení šíření fámy, který simuluje fakt, že šiřitel může fámu zapomenout anebo sám od sebe ztratí zájem ji šířit bez potřeby kontaktu s někým jiným. *Šiřitel* tedy může sám od sebe přestat šířit fámu a s pravděpodobností se z něj stává *potlačovatel*.

Rozšířený model funguje takto:

* Pokud přijde *neznalý* do kontaktu s *šiřitelem*, *neznalý* se okamžitě “nakazí“ a stane se z něj *šiřitel* s pravděpodobností . Tento krok funguje identicky i v Maki – Thompson modelu.
* *Šiřitel* spontánně přestává šířit fámu s pravděpodobností
* Pokud se *šiřitel* potká s jiným *šiřitelem* anebo *potlačovatelem,* zjistí, že informace je již známá, ztratí zájem o její další šíření a z iniciujícího *šiřitele* se stane *potlačovatel* spravděpodobností. Tento krok funguje identicky i v Maki-Thompson modelu.

Autoři aplikovali tento model na různé typy sítí, například náhodnou síť vygenerovanou Erdős-Rényi algoritmem [15] anebo bez-škálovou síť se stupňovým exponentem . Jelikož distribuce vrcholů náhodné sítě se řídí Poissonovo rozdělením, a ne mocninným zákonem, nemá tak velké centrální vrcholy. [15] Tato vlastnost způsobuje, že náhodné sítě jsou mnohem méně náchylné na šíření nemocí i fám, což zmiňují i sami autoři studie.

----------------------------------------------------------assfdssd

Model se zaměřuje spíše na všeobecné šíření fám v síti než konkrétně na řetězové emaily, ale je možné jej aplikovat na jakoukoliv všeobecnou síť. Různé konspirační teorie, či hoaxy lze považovat za fámu. Z těchto důvodů předpokládám, že při jeho aplikaci na konkrétní emailovou síť by bylo možné aproximovat i propagaci řetězového emailu s takovýmto obsahem. Na základě tohoto předpokladu jsem implementoval lehce modifikovanou variantu modelu, kterou jsem aplikoval na konkrétní emailovou síť rozšířenou o průměrné reakce na konspirační narativy. Tyto reakce byly odvozeny z výsledků studie společnost nedůvěry [25], která se zabývá dopady konspirací, dezinformací a všeobecné nedůvěry v instituce na českou společnost.

## 3.2 Liben-Nowell a Kleinberg model šíření (LNK model)

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 petice pro zachování vládní finanční podpory pro veřejnoprávní rádiovou stanici NPR a veřejnoprávní televizní stanici PBS ve Spojených Státech z roku 1995. Data sesbírali z veřejných internetových archivů. [1]

Za normálních okolností je téměř nemožné legálně vystopovat cestu řetězového emailu. Jelikož na rozdíl od sociálních sítí emailové zprávy jsou soukromé a buď příjemce anebo odesílatel emailu se musí sami 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.

### 3.2.1 Rekonstrukce šíření petice z reálných dat [1,28]

### Autoři v první fázi výzkumu hledali kopie petic pomocí různých internetových vyhledávačů. Pro petici proti válce v Iráku autoři petice použili jako vyhledávací parametr použili jméno a bydliště prvního signatáře a pro petici na podporu veřejnoprávních médií použili text obsahu petice. U petice proti válce v Iráku takto získali 1265 unikátních webových stránek. Nalezené webové stránky autoři prozkoumali a vyřadili všechny stránky, které neobsahovaly kopii petice. Mezi vyřazenými stránkami byly například různé pojednání o petici anebo rozhovory s jejím prvním signatářem. 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 odstranili a všechna jména příjemců převedli na malá písmena. Následně si 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. Pro každý identifikátor provedli kroky 3 a 4. Po zpětné ruční kontrole autoři zjistili, že všechny kopie petice proti válce v Iráku 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

### Výše uvedený postup také odhalil duplicity v kopiích petice a díky zavedení parametru editační vzdálenosti pomohl unifikovat kopie, které se od sebe lišily pouze typograficky. Po ruční kontrole autoři zjistili, že pokud je editační vzdálenost nastavena na hodnotu 6, tak všeobecně nedocházelo k záměně dvou různých jmen. Stejný postup byl aplikován u petice pro zachování vládní podpory pro veřejnoprávní média NPR a PBS.

### Zpracování reálných dat ukázalo, že autoři získali 637 unikátních kopií petice proti válce v Iráku s podpisy 18 119 lidí a 316 unikátních kopií s 13 052 podpisy pro podporu NPR a PBS.

Struktury vytvořené z dat zkoumaných petic pomocí výše uvedeného postupu jsou stromy s nízkým počtem rozvětvení a vysokou hloubkou. Hloubka vrcholu je definována jako jeho vzdálenost od kořene stromu.

Strom rekonstruovaný z protiválečné petice měl 18 119 vrcholů z toho 94.26 % mělo pouze jednoho potomka, medián hloubky všech vrcholů byl 288 a šířka stromu byla 82. Šířku stromu autoři definují jako maximální počet vrcholů, které mají stejnou hloubku. Z petice na podporu NPR a PBS bylo rekonstruováno několik stromů, kvůli lehce odlišnému způsobu získání dat. Vyhledávání neprobíhalo pomocí prvního signatáře, ale pomocí textu petice, což mělo za důsledek více prvotních signatářů, tedy i více stromů. Stromů bylo celkem 12, z toho 3 největší měly několik tisíc vrcholů, velikost dalších dvou se pohybovala okolo jednoho tisíce a zbytek stromů měl méně než tisíc vrcholů. Autoři uvádí podrobnější parametry pouze pro tři největší stromy.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Strom 1 | Strom 2 | Strom 3 |
| Počet vrcholů | 3032 | 2442 | 3250 |
| Počet vrcholů s přesně 1 potomkem | 2870 (94.66 %) | 2326 (95.25 %) | 3117 (95.91 %) |
| Medián hloubky vrcholů | 221 | 269.5 | 231 |
| Šířka stromu | 16 | 11 | 17 |

*Tabulka 1: parametry tří největších stromů rekonstruovaných z NPR petice [28]*

Tyto rekonstruované stromy neodpovídají výsledkům ostatních modelů založených na epidemiologickém šíření, kde i při větším počtů vrcholů jsou vzdálenosti ve výsledném grafu krátké a vrcholy mají typicky větší množství sousedů, než je tomu u těchto vygenerovaných stromů. [1,11]

Tento přístup popisující šíření petic má tu výhodu, že se zakládá na 100 % reálných datech. Nevýhodou je, že tato data nejsou úplná, jelikož z principu nelze zachytit všechny instance přeposlání zprávy.

Obsah obrázku skica, kresba, umění, černobílá

Popis byl vytvořen automaticky

*Vizualizace zrekonstruovaného stromu šíření petice proti válce v Iráku [28]*

Obsah obrázku skica, kresba

Popis byl vytvořen automaticky

*Přiblížený pohled na část zrekonstruovaného stromu šíření petice proti válce v Iráku [1]*

### Model šíření petic

### Na základě zrekonstruovaných dat šíření autoři navrhli model, který by popisoval šíření petic lépe než modely odvozené od epidemiologických modelů. Tento model se odlišuje od ostatních modelů [11] tím, že je odvozen od šíření konkrétních emailů.

### Vzhledem k tomu, že velké reálné sítě, včetně sociálních a emailových, jsou většinou bez-škálové [13,15,20], šíření v nich probíhá expanzivně a v malém počtu kroků, ale s velkým dosahem. Toto je způsobeno existencí centrálních vrcholů s extrémně vysokým počtem sousedů, které při zapojení se do šíření distribuují zprávu k velkému množství vrcholů. 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. Tato struktura neodpovídá ostatním modelům, kde průměrné vzdálenosti mezi vrcholy jsou malé.

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 rozsahu od 0.5 do 0.75 nemá zásadní vliv na výslednou strukturu šíření. [1]

Pokud se příjemce rozhodne reagovat, tak má dvě vzájemně se vylučující možnosti reakce. První a častější variantou, je že na email 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 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 ji 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 snazší objekt ke zkoumání než jiné typy emailů.

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 vrcholy 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.

Existují také tzv. super-šiřitelé, kteří jsou schopni zaslat email na řádově stovky adres. Tento fakt je reprezentován centrálními vrcholy v síti, které se rozhodly zprávu přeposlat. [21]

Autoři pro simulaci nepoužili emailovou síť, ale reprezentaci několika velkých sociálních sítí, z nichž největší byla sociální síť LiveJournal, která obsahovala 4.4 milionu vrcholů a jejíž vlastnosti jsou podle autorů srovnatelné s ostatními velkými sociálními sítěmi. Dalšími sítěmi, které byly použity pro simulaci jsou komunikační síť editorů Wikipedie a síť spoluautorství vytvořená z bibliografické databáze DBLP. O těchto sítích autoři neposkytují přesnější informace, kromě toho, že obě mají přibližně desetinu vrcholů jako síť LiveJournal. Výsledky modelu aplikovaného na tyto sítě se od sebe zásadně nelišily co do struktury výsledků, tedy opět se jednalo o stromy s vysokou hloubkou a většina vrcholů měla pouze jednoho potomka. Jediným zásadním rozdílem byl počet vrcholů simulovaných stromů, což souvisí s velikostí sítí, nad kterými byl model aplikován, čím větší síť, tím větší potenciální dosah zprávy. Přesná data o konkrétních výsledcích simulace nad jednotlivými sítěmi autoři neuvádí.

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 výsledků šíření ostatních modelů ani vzdáleně nepřipomíná strom s vysokou hloubkou a malou šířkou.

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 . Realistický 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.

Další vlastností tohoto modelu, která ovlivňuje výslednou strukturu, je fakt, že tento model nepopisuje dosah zprávy, ale pouze mapuje její “viditelnou” část, tedy tu, získanou na základě zveřejnění zprávy. Tato část je menší než kompletní mapa šiřitelů, protože ne každý zprávu zveřejní.

## 3.3 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 jednoduchý postup pro generování stromů, který byl publikován již v 19. století.

V návaznosti na studii *Tracing information flow on a global scale using Internet chain-letter data*, ve které popisují David Liben-Nowell a Jon Kleinberg principy šíření internetových petic, vznikla studie *Using selection bias to explain the observed structure of Internet diffusions.* Autoři této návazné studie popisují alternativní způsob simulace výsledků LNK modelu, kterým je právě Galton – Watsonův algoritmus pro generování stromů.

Algoritmus funguje následujícím způsobem [5]:

1. Založí graf s jedním vrcholem, který je označen jako aktuální rodič
2. Založí prázdnou frontu a nastaví pořadí aktuální generace na 0
3. Podle předem dané pravděpodobnostní distribuce náhodně vygeneruje počet potomků pro aktuálního rodiče
4. Vytvoří množinu vrcholů o velikosti . Pro každý vrchol vykoná následující kroky:
   1. Vloží do .
   2. Vytvoří hranu mezi vrcholy a .
   3. Vloží vrchol do fronty
5. K pořadí aktuální generace  přičte 1
6. Pokud je fronta prázdná nebo je větší nebo rovna předem určenému limitu počtu generací , algoritmus skončí. V opačném případě vykoná následující kroky:
   1. Vyjme první vrchol z fronty a označí jej jako aktuálního rodiče
   2. Algoritmus vykoná kroky 3-6 pro nového aktuálního rodiče .

Mezi vstupními parametry figurují distribuce potomků a limit počtu generací. Distribuce určuje pro celočíselný počet potomků jeho pravděpodobnost výskytu . Bez nastavení limitu maximálního počtu generací by algoritmus generoval nekonečný strom, v případě, že .

Pro jeden ze stromů rekonstruovaný z petice na podporu veřejnoprávního vysílání, který byl použit pro odvození LNK modelu, odvodili autoři návazné studie distribuci uvedenou v tabulce 2. Pravděpodobnost je zde jednoduše odvozena od procentuálního zastoupení počtu potomků. [4]

|  |  |
| --- | --- |
| k | p(k) |
| 0 | 0.0246 |
| 1 | 0.9525 |
| 2 | 0.0217 |
| 3 | 0.0012 |
|  | 0 |

*Tabulka 2: pravděpodobnost počtu potomků pro simulaci stromu rekonstruovaného z petice na podporu veřejnoprávního vysílání [4]*

I když je výše uvedená distribuce odvozena od konkrétního stromu (stromu 2 v tabulce 1), lze s její pomocí relativně přesně modelovat všeobecné výsledky LNK algoritmu, jelikož distribuce vrcholů mezi stromy jednotlivých petic se od sebe řádově liší v desetinách až nízkých jednotkách procent. [1, 28]

Autoři nasimulovali celkem 10 000 stromů pomocí Galton-Watsonova modelu, kde jako vstupní parametr použili distribuci potomků v tabulce 2. Maximální limit počtu generací ponechali bez omezení, jelikož pravděpodobnost , což zajišťuje, že generování stromu nebude probíhat nekonečně dlouho.

Z těchto 10 000 vygenerovaných stromů jich pouze 654 mělo více než 2442 vrcholů a pouze 110 z nich mělo počet vrcholů v rozmezí 2442 až 3250. Hodnoty určující toto rozmezí odpovídají velikosti stromů rekonstruovaných z petice na podporu veřejnoprávního vysílání uvedené v tabulce 1. Tyto výsledky poukazují na významný fakt, že většina petic či všeobecně řetězových emailů se nedostane k velkému počtu uživatelů. [4]

Autoři u těchto vygenerovaných stromů pozorovali také medián hloubky jednotlivých vrcholů a šířku stromu. Šířka stromu je, stejně jako u LNK modelu, definována jako maximální počet vrcholů, které mají stejnou hloubku. Pro tyto pozorované vlastnosti sesbírali statistiky kumulovaně pro všechny vygenerované stromy a pro stromy spadající do rozmezí od 2442 do 3250 co do počtu vrcholů.

Obsah obrázku text, diagram, snímek obrazovky

Popis byl vytvořen automaticky

*Obrázek 8: Statistiky simulovaných stromů. V levém sloupci jsou statistiky pro všechny simulované stromy, v pravém jsou statistiky pro stromy jejichž počet vrcholů se pohyboval v rozmezí od 2442 do 3250. Bílé kruhy v grafech reprezentují reálné stromy zrekonstruované z petice na podporu NPR. [4]*

# Implementace modelů šíření a analýza jejich výsledků

## 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-Albertové algoritmu. [16]

V rámci této práce používám pouze neorientované grafy, veškeré orientované grafy jsem tedy převedl na neorientované. Toto jsem učinil z toho důvodu, že pro potřeby této práce jsem potřeboval všeobecnou emailovou síť, co nejméně závislou na původní komunikaci a směr původního emailu, který reprezentuje hranu, je tedy v tomto případě irelevantní.

U emailových sítí jsem se zaměřil na data z velké výzkumné evropské organizace. Data jsem stáhnul z veřejně přístupné databáze grafů Stanfordské univerzity (SNAP). [2,3,12] O jakou konkrétní organizaci se jedná není u zdroje dat specifikováno, ale pro účely této práce to není důležité, protože se v ní zabývám všeobecnými emailovými sítěmi a nezkoumám konkrétní řetězový email anebo síť, po které se šířil. Anonymizovaná data pochází z října 2003 až května 2005. Z přibližně 3 milionů emailů byly sestaveny dva grafy, kde každý vrchol reprezentuje unikátní emailovou adresu a každá orientovaná hrana mezi adresami reprezentuje alespoň jeden vyměněný email.

První graf rekonstruovaný z komunikace výzkumné instituce (dále jen *EuAll*) reprezentuje všechny emailové adresy, které se zúčastnily komunikace s někým z výzkumné organizace. Druhý takto sestavený graf (dále jen *EuCore*) byl vytvořen pouze z “jádra” emailové sítě, tedy z těch adres, které náležely přímo výzkumné organizaci a z interních emailů, které si mezi sebou zaměstnanci instituce posílali.

Graf *EuAll* má následující charakteristiky. Obsahuje 265214 vrcholů a 420045 orientovaných hran, ze kterých jsem převedením získal 365572 neorientovaných hran. Průměrný shlukovací koeficient vrcholů v grafu je 0.067. Průměr grafu je 14. Průměr je definovaný jako největší minimální délka cesty mezi všemi dvojicemi vrcholů v grafu.

Graf *EuCore* obsahuje 1005 vrcholů a 25571 orientovaných hran, ze kterých jsem převedením získal graf s 16706 neorientovanými hran. Průměrný shlukovací koeficient vrcholů je 0.4 a průměr grafu je 7.

Core není scalefree

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.

* 1. **Popis implementovaného nástroje pro modelování a vizualizaci šíření**

**4.3 Implementace Galton – Watson modelu a jeho výsledky**

Motivací pro implementaci tohoto modelu bylo získání referenčních dat pro LNK model. Jelikož jsem se rozhodl modifikovat LNK model, tak aby lépe odpovídal všeobecnému šíření řetězových emailů, a ne pouze peticím, které obsahují historii podpisů, potřeboval jsem data, se kterými by bylo možné výsledky modifikovaného LNK modelu porovnat. Autoři studie *Using selection bias to explain the observed structure of internet diffusions* ukázali, že výsledky LNK modelu lze velmi věrohodně napodobit právě Galton-Watson modelem s vhodně nastavenými vstupními parametry. [4]

**4.3.1 Implementace modelu**

Pro implementaci tohoto modelu jsem zvolil postup popsaný v kapitole 3.3. Jelikož jsem se co nejvíce snažil přiblížit výsledkům výše zmíněné studie porovnávající LNK model s Galton-Watson modelem, použil jsem stejnou vstupní distribuci počtu potomků jako její autoři, tedy tu uvedenou v tabulce 2. Jako limit počtu generací jsem použil hodnotu 10 000. Nastavení limitu jsem učinil, protože výjimečně se stává, že algoritmus bez tohoto omezení generuje obrovské stromy, řádově o velikosti vyšších stovek tisíc či několika milionů vrcholů. Stromy o této velikosti výrazně zpomalovaly generování většího množství stromů zároveň a jejich následnou analýzu. Limit 10 000 zároveň nezabraňuje vzniku stromů o velikosti typické pro pozorované petice, tedy řádově o velikosti několika tisíc vrcholů.

Pro sbírání statistik o vlastnostech stromů jsem použil algoritmus BFS, kterým jsem prošel každý strom a při průchodu zaznamenával statistiky.

**4.3.2 Výsledky a jejich analýza**

Vygeneroval jsem celkem 10 000 stromů, stejné množství jako autoři studie porovnávající LNK a Galton-Watson modely. Sledoval jsem u nich stejné metriky jako autoři, tedy počet vrcholů, medián hloubky vrcholů a jejich šířku.

Z těchto 10 000 stromů jich 599 mělo více jak 2442 vrcholů a 109 jich mělo počet vrcholů v rozmezí od 2442 do 3250. Tyto hodnoty jsou velmi podobné výsledkům, které získali autoři studie, v jejich simulaci mělo 654 stromů více než 2442 vrcholů a 110 se jich pohybovalo v rozmezí od 2442 do 3250, co do počtu vrcholů. Rozdíly ve výsledcích jsou v rámci statistické chyby.

Rozmezí 2442 až 3250 reprezentuje velikost nejmenšího a největšího stromu zrekonstruovaného z NPR petice a z jejich distribuce počtu potomků byl odvozen tento vstupní parametr. Parametry těchto stromů pozorují i autoři studie.

Stromy, které jsem vygeneroval svojí implementací mají vlastnosti velmi podobné stromům vygenerovaným autory studie. Autoři neposkytli přesná čísla, pouze histogramy orientačně znázorňující výsledky. Pro srovnání jsem vytvořil podobné histogramy popisující vlastnosti mnou vygenerovaných stromů.

Již na první pohled je patrné, že naprostá většina stromů měla relativně malou hloubku i šířku. Konkrétně 8962 stromů mělo medián hloubky všech vrcholů menší než 150 a 9453 stromů mělo šířku menší než 15.

Tyto výsledky odpovídají faktu, že většina (94 %) stromů nedosáhla ani na spodní limit 2442 vrcholů. Tento fakt odpovídá důležité skutečnosti, že většina emailů má relativně malý dosah a jejich šíření ve velkém měřítku není nemožné, ale je velmi nepravděpodobné. Najdou se však i výjimky, největší strom měl 651 518 uzlů, šířku 277 a medián hloubky vrcholů 5319 a to narazil na limit počtu generací.

Obsah obrázku text, snímek obrazovky, displej, číslo

Popis byl vytvořen automatickyPři testování algoritmu bez tohoto limitu jsem narazil i na stromy s několika miliony vrcholů. Toto poukazuje na možnost existence řetězových emailů s obrovským dosahem, které jsou sice velmi vzácné, ale jejich existence není vyloučena. Je však nutné zdůraznit, že LNK model popisuje pouze pozorovatelnou část řetězových emailů a je na rozdíl od Galton-Watson modelu limitován sítí nad kterou je spouštěn. Pokud by tedy existoval řetězový email, jehož pozorovatelná část by měla velikost v řádově vyšších stovkách tisíc vrcholů, jeho skutečný dosah by byl ještě několikanásobně vyšší, tedy klidně i v řádech desítek milionů. U takovýchto emailů však již narážíme na limitace reálného světa, jako je například celkový počet uživatelů emailu, který rozumí jazyku, v němž je napsán. Například email napsaný v češtině je přibližně limitován velikostí populace České, případně Slovenské republiky, která čítá zhruba 15 milionů, a to opomíjím fakt, že ne všichni lidé používají email.

*Obrázek 9: Histogram mediánu hloubek vrcholů všech stromů. Vlastní zpracování*

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

Popis byl vytvořen automaticky

*Obrázek 10: Histogram šířek všech stromů. Vlastní zpracování*

Při zaměření se na stromy v limitu reálných pozorovaných emailů (2442-3250), je z histogramů patrné, že moje simulace opět vykazovala podobné vlastnosti jako simulace autorů studie. Byť zde jsou rozdíly o něco větší, například počet stromů s šířkou 14 byl výrazně nižší, než počet stromů s šířkou 15. U těchto nesrovnalostí předpokládám, že byly způsobené malým vzorkem dat, histogramy byly vytvářeny ze 109 stromů splňující limity počtu vrcholů.

Základní parametry, tedy že velká část stromů splňující limity má šířku v rozpětí 13 až 15, medián hloubky uzlů se často pohyboval mezi hodnotami 200 až 300 a žádný strom neměl medián hloubky uzlů menší než 100 byly zachovány. Konkrétně 40 stromů mělo šířku v rozpětí 13 až 15 a 54 jich mělo medián hloubky uzlů v rozpětí 200 až 300.

Na základě analýzy výsledků modelu, je tedy možné prohlásit, že cíl naimplementovat vhodný referenční model pro LNK byl splněn.

Obsah obrázku text, snímek obrazovky, diagram, Písmo

Popis byl vytvořen automaticky

*Obrázek 11: Histogram mediánu vrcholů stromů s počtem vrcholů mezi 2442 a 3250. Vlastní zpracování*

Obsah obrázku text, snímek obrazovky, diagram, Vykreslený graf

Popis byl vytvořen automaticky

*Obrázek 12: Histogram šířek stromů s počtem vrcholů mezi 2442 a 3250. Vlastní zpracování*

Obsah obrázku text, diagram, snímek obrazovky

Popis byl vytvořen automaticky

*Obrázek 8: Referenční výsledky autorů studie [4]*

Obsah obrázku vzor, červená, design, umění

Popis byl vytvořen automaticky

*Obrázek 13: přiblížený pohled na vygenerovaný strom s 2756 uzly a šířkou 15. Vlastní zpracování*

## 4.4 Implementace modifikovaného LNK modelu a jeho výsledky

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 modifikovanou variantu u které se snažím lépe napodobit chování všeobecného řetězového emailu.

**4.4.1 Implementace modelu**

Modifikovaný model se stejně jako jeho originální varianta navržená autory studie *Tracing information flow on a global scale using Internet chain-letter data* a kterou popisuji v kapitole 3.2, spouští, nad již existujícími sítěmi. Pro potřeby tohoto algoritmu jsem využil sítě *EuAll* a *EuCore,* popsané v kapitole 4.1.

Na začátku běhu si algoritmus vybere počáteční vrchol, což může proběhnout jedním ze dvou způsobů, podle konfigurace konkrétního běhu. Prvním způsobem je stejně jako v originální variantě výběr náhodného vrcholu v grafu, čímž se snažím reprezentovat spontánní vzniknutí zprávy, kde kdokoliv z jakéhokoliv důvodu začne šířit obsah, který mu přijde zajímavý. Druhým způsobem je start z největšího centrálního vrcholu, který simuluje záměrné šíření s cílem dostat zprávu k co největšímu počtu lidí. Tento způsob je typický pro šíření nepravdivé propagandy cizích států.

Vybraný počáteční vrchol následně rozešle zprávu všem svým sousedům. Sousedi jsou následně označeni jako příjemci a na zprávu s určitým zpožděním reagují. Zpoždění je určené distribucí času odpovědi podle distribuční funkce , převzaté z originálního modelu. Čas neplyne kontinuálně, ale ve velmi malých diskrétních krocích, které toto plynutí simulují. Krok zpoždění je reprezentován intervalem v distribuční funkci, jeho zvolení probíhá náhodně podle přiřazených vah. Váha kroku je dána určitým integrálem intervalu . Hodnota kroku zpoždění reprezentuje počet kroků, po které příjemce zprávy nic nedělá. Příjemce reaguje na zprávu v čase , kde je počet časových kroků od začátku běhu algoritmu v čase obdržení zprávy a je zvoleno náhodně podle vygenerovaných vah kroků.

Na začátku běhu algoritmus se nastaví aktuální čas . Poté postupuje po krocích a v každém kroku aktivuje vrcholy, pokud je jejich čas reakce . Po aktivování všech vrcholů splňující tuto podmínku, zvýší hodnotu aktuálního času o 1 ().

Každý příjemce zprávy v čase reaguje na zprávu, tato reakce se řídí následujícími pravidly:

1. Příjemce se z pravděpodobností rozhodne se na zprávu reagovat v kroku 2. V opačném případě se rozhodne zprávu ignorovat a žádný další krok se pro tento vrchol nevykoná.
2. Pokud se příjemce rozhodl zprávu neignorovat pak se z něj s pravděpodobností stává odesílatel a nastávají následující kroky:
   1. Odesílatel zprávu přepošle všem svým sousedům, kteří nejsou označeni jako příjemci.
   2. Všichni sousedé, kteří takto obdrželi zprávu jsou označeni jako příjemci a je jim přiřazen aktivační čas , kde a je separátně vygenerováno pro každého nového příjemce pomocí váženého náhodného výběru popsaného výše.
   3. Odesílatel je označen jako veřejný odesílatel s pravděpodobností .
3. Pokud se z příjemce nestal odesílatel, stává se z něj skupinový respondent s pravděpodobností .

Označení vrcholu jako příjemce mu zůstává po celou dobu běhu algoritmu nezávisle na jeho dalších označeních, které může v průběhu získat. Tento mechanismus zaručuje, že každý vrchol, který zprávu dostane na ni reaguje pouze jednou, stejně jako v originálním modelu.

Pravděpodobnosti ignorování zprávy a její hodnota 0.65 jsou převzaty z originálního modelu. Toto jsem učinil, protože předpokládám, že ignorování zprávy je závislé na tom, jestli se jedná o petici anebo jakýkoliv řetězový email.

Beze zásadní změny jsem ponechal i hodnotu , pravděpodobnost zveřejnění emailu. Jednou ze zajímavých vlastností originálního modelu, je fakt, že nesimuluje absolutní dosah zprávy, ale pouze její pozorovatelnou část. Hodnotu jsem ponechal stejnou jako v originále tedy 0.22. U petice se dá předpokládat, že se ji lidé primárně snaží šířit z dobré vůle, jelikož upřímně věří jejímu obsahu, a proto ji zveřejňují jako sekundární druh podpory. U motivace zveřejnění řetězového emailu očekávám, že 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 podobný druh motivace pro zveřejnění, typický pro lidi s proti-systémovým smýšlením, kteří mají tendenci obsahu emailu věřit a snaží se o něm zvýšit také zvýšit povědomí, jelikož obsažené informace mohou vnímat jako pravdu utajovanou státními institucemi a mainstreamovými médii.

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.

Hodnoty této pravděpodobnosti jsem použil relativně podobné jako autoři LNK modelu, konkrétně v intervalu od 0.83 do 0.95. Skupinová odpověď ovšem nemá v tomto případě efekt vytváření dlouhých řetězců propagace, protože díky absenci podpisů, nemůže jméno respondenta 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 procenta příjemců [24].

Po dokončení simulace se z výsledků vytváří několik grafů, reprezentující různé pohledy na šíření. V každém výsledném grafu vrchol reprezentuje uživatele, který na email určitým způsobem reagoval a hrana mezi dvěma vrcholy reprezentuje, že mezi nimi simulovaný řetězový email byl přeposlán. Podle výše uvedeného postupu algoritmus označoval vrcholy jako odesílatele, veřejného odesílatele a skupinového respondenta. Na základě těchto označení jsou vytvořeny výsledné grafy.

První graf, dále jen graf odesílatelů obsahuje všechny odesílatele i ty veřejné, neobsahuje však skupinové respondenty. Tento pohled reprezentuje úplnou mapu šiřitelů konkrétního emailu.

Druhý graf, dále jen viditelný graf odesílatelů,

asdasdasdasd

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.

**4.4.2 Výsledky a jejich analýza**

**4.5** **Implementace všeobecného modelu šíření fám a jeho výsledky**

# 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. Barabási, A.-L., & Pósfai, M. (2017). Chapter 1 Introduction. In *Network science*. essay, Cambridge University Press. Dostupné také z: <http://networksciencebook.com/chapter/1>
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*, 107(24),10833–10834. Dostupné z: <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., 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.* Dostupné z: <https://doi.org/10.1016/j.physa.2006.07.017>
12. Zdroj eu core
13. Ebel, H., Mielsch, L.I., & Bornholdt, S. (2002). Scale-free topology of e-mail networks*. Phys. Rev. E, 66, 035103.* Dostupné z: <https://doi.org/10.1103/PhysRevE.66.035103>
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. <https://www.irozhlas.cz/zpravy-domov/spolecnost-neduvery-serial-dezinformace-vyzkum-skupiny-seznam_2306120500_pik>
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>

|  |  |  |
| --- | --- | --- |
| **Seznam obrázků** | 11 |  |
| Obrázek 1 - Části CI/CD .......................................................................................................................................................... |  |
| Obrázek 2 - Proces Continuous Integration ................................................................................................................... | 12 |  |
| Obrázek 3 - Popis důvodů konkurenčních výhod ........................................................................................................ | 15 |  |
| Obrázek 4 - Rozdíl mezi Continuous Delivery a Deployment................................................................................. | 15 |  |
| Obrázek 5 - Lokální správa verzí. ....................................................................................................................................... | 16 |  |
|  |  |  |
|  |  |  |