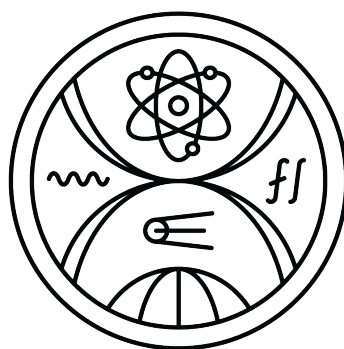


UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



SENTIMENT ANALÝZA A MODELOVANIE TÉM
SPRAVODAJSKÝCH ČLÁNKOV POMOCOU TEÓRIE
KOMPLEXNÝCH SIETÍ

DIPLOMOVÁ PRÁCA

2024

BC. JÁN MAJERSKÝ

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

SENTIMENT ANALÝZA A MODELOVANIE TÉM
SPRAVODAJSKÝCH ČLÁNKOV POMOCOU TEÓRIE
KOMPLEXNÝCH SIETÍ
DIPLOMOVÁ PRÁCA

Študijný program: Aplikovaná informatika
Študijný odbor: Informatika
Školiace pracovisko: Katedra aplikovanej informatiky
Školiteľ: doc. RNDr. Damas Gruska, PhD.

Bratislava, 2024
Bc. Ján Majerský

ZADANIE ZÁVEREČNEJ PRÁCE

[download this from AIS]

PodĎakovanie: Chcel by som sa poĎakovať svojmu školiťovi doc. RNDr. Damasovi Gruskovi, PhD. za spoluprácu, možnosť nahliadnúť do sveta komplexných sietí a príležitosť pokračovať v predošlej práci.

Abstrakt

Svetové spravodajstvo odzrkadľuje aktuálne dianie vo svete a spoločnosti v ktorej žijeme. Vďaka jeho internetovej podobe majú dáta o spravodajských článkoch hojnú kvantitu a pohodlnú dostupnosť. Ich analýzou vieme identifikovať (najfrekvencovanejšie) pojmy a témy ktoré sa spájajú s rôznymi zaujímavými oblasťami, vyhodnotiť sentiment s ktorým sa dané pojmy spomínajú a sledovať vývoj týchto atribútov v čase. Cieľom tejto práce je zamerať sa na kľúčové slová v spravodajských článkoch. Kľúčové slová modelujeme do komplexných sietí a pomocou teórie komplexných sietí ich následne analyzujeme. Navrhujeme postupy pre analýzu sentimentu, modelovanie tém a ich vizualizáciu, ktoré sú použiteľné pre ľubovoľnú množinu článkov. V rámci implementačnej časti práce vznikol aj softvérový nástroj na ktorom naše postupy demonštrujeme.

Kľúčové slová: complex networks, news articles, keywords, sentiment, topic modeling

Abstract

World news reflects current events in the world and the society we live in. Thanks to its web-based format, data about news articles is abundant in quantity and conveniently accessible. By analyzing them, we can identify the (most frequent) terms and topics that are associated with different areas of interest, evaluate the sentiment with which the terms are mentioned, and track the evolution of these attributes over time. The aim of this thesis is to focus on keywords in news articles. We model the keywords into complex networks and then analyze them using complex network theory. We propose procedures for sentiment analysis, topic modeling, and visualization which are applicable to any set of articles. In the implementation part of the thesis, we also develop a software tool on which we demonstrate our procedures.

Keywords: English, keywords, here

Obsah

Úvod	1
1 Teoretické východiská	3
1.1 Spracovanie prirodzeného jazyka	3
1.1.1 Lematizácia	3
1.1.2 Extrakcia kľúčových slov	3
1.1.3 Sumarizácia	3
1.1.4 Analýza sentimentu	3
1.1.5 Modelovanie tém	3
1.2 Teória grafov a komplexných sietí	3
1.2.1 Graf	4
1.2.2 Jednoduchý graf	4
1.2.3 Ohodnotený graf	5
1.2.4 Matica susednosti	5
1.2.5 Kompletný graf	6
1.2.6 Podgraf	7
1.2.7 Klika	7
1.2.8 Cesta	7
1.2.9 Komponent	8
1.2.10 Priemerná najkratšia dĺžka cesty	8
1.2.11 Stupeň vrcholu	9
1.2.12 Distribúcia stupňov	9
1.2.13 Klasterizačný koeficient	10
1.2.14 Náhodné grafy	10
1.2.15 Siete malého sveta	11
1.2.16 Bezškálové siete	12
1.2.17 Komplexné siete	13
1.2.18 Detekcia komunit	14
1.2.19 Centrálnosť blízkosti	15

2	Vybrané postupy pre zber dát, analýzu a vizualizáciu	17
2.1	Zber dát	17
2.2	Extrakcia kľúčových slov	17
2.3	Analýza sentimentu	17
2.4	Modelovanie komplexných sietí	17
2.5	Analýza a modelovanie tém	17
2.6	Vizualizácia	17
2.7	Limity navrhnutých postupov	17
3	Softvérový nástroj	19
3.1	Použité technológie	19
3.2	Implementácia	19
4	Výskum výstupov	21
	Záver	23

Zoznam obrázkov

1.1	Známy príklad sociálnej siete (grafu) nazývanej <i>Zachary's Karate Club</i> (štandardne používaný príklad).	4
1.2	Porovnanie jednoduchého grafu (naľavo) a grafu so slučkami a viacnásobnými hranami (napravo).	5
1.3	Matica susednosti jednoduchého neohodnoteného grafu.	6
1.4	Kompletné grafy s rôznym počtom vrcholov.	6
1.5	Klika (čierna farba) v neorientovanom grafe, ktorá obsahuje 4 vrcholy. .	7
1.6	Cesta (čierna farba) medzi dvoma vrcholmi, ktorá je zároveň aj ich najkratšia možná.	8
1.7	Nesúvislý graf s dvoma komponentmi A a B.	8
1.8	Vizualizácia distribúcie stupňov <i>Zachary's Karate Club</i> siete (Obr. 1.1). .	10
1.9	Náhodné grafy s 10 vrcholmi $G(10, p)$ s pravdepodobnosťami p v intervale $[0, 1]$	11
1.10	Sieť malého sveta (naľavo) a náhodný graf (napravo) s rovnakým množstvom vrcholov (10) a hrán (v priemere 4 hrany na vrchol).	11
1.11	Porovnanie bezškálovej siete s hubmi (naľavo) a náhodného grafu (napravo). .	12
1.12	Mocninová distribúcia stupňov (naľavo) a binomická distribúcia stupňov (napravo).	13
1.13	Komunity (šedé oblasti) v malom grafe.	15

Úvod

You can reference other chapters, sections, images, etc. using the `cleveref` package like this: chapter 1.

L^AT_EX basics tutorial: https://www.overleaf.com/learn/latex/Learn_LaTeX_in_30_minutes

Kapitola 1

Teoretické východiská

Vo východiskovej kapitole popisujeme a vysvetľujeme teóriu, potrebnú pre úplne pochopenie postupov tejto práce. V prvej sekcii sa venujeme vybraným oblastiam spracovania prirodzeného jazyka. V druhej sekcii sú to základné pojmy z oblasti teórie grafov a komplexných sietí, ktoré súvisia s touto prácou. Táto sekcia je z veľkej časti prebraná z bakalárskej práce, pre prehľadnosť ju však uvádzame aj tu.

1.1 Spracovanie prirodzeného jazyka

1.1.1 Lematizácia

1.1.2 Extrakcia kľúčových slov

1.1.3 Sumarizácia

1.1.4 Analýza sentimentu

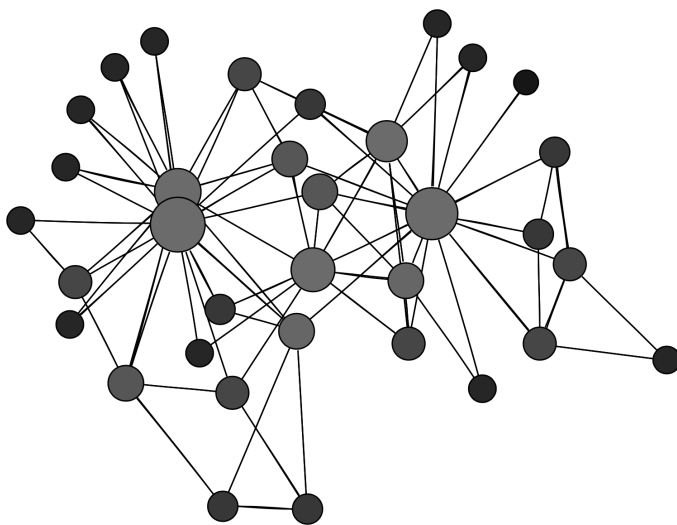
1.1.5 Modelovanie tém

1.2 Teória grafov a komplexných sietí

Štúdium komplexných sietí je interdisciplinárny odbor, ktorý spája myšlienky z rôznych vedných oblastí ako matematika, fyzika, informatika, biológia, sociálne vedy a mnohé ďalšie [1]. Keďže komplexné siete sú vo svojej podstate matematické grafy [1, 2], teória komplexných sietí je príbuzná teórii grafov a môžeme ju považovať za akúsi jej nadstavbu.

1.2.1 Graf

Graf (môžeme použiť aj označenie sietí) $G = (V, E)$ je definovaný ako usporiadaná dvojica dvoch množín. Konečnou neprázdnu množinou vrcholov V a množinou hrán E , kde E je množina dvojprvkových podmnožín množiny V . Jednoducho povedané, hrany spájajú vrcholy (Obr. 1.1). Dva koncové vrcholy tej istej hrany voláme *susedné vrcholy* [3].

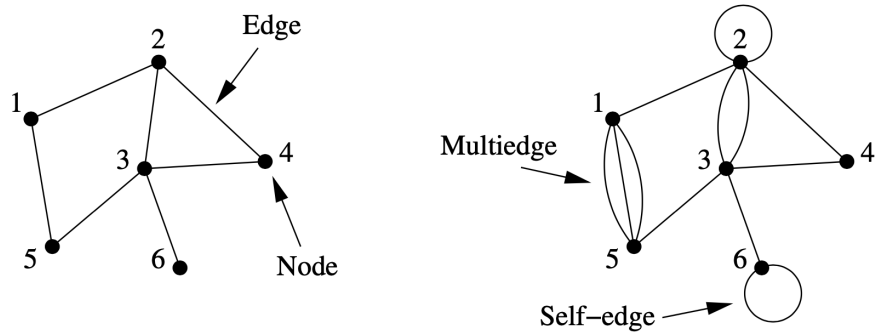


Obr. 1.1: Známy príklad sociálnej siete (grafu) nazývanej *Zachary's Karate Club* (štandardne používaný príklad).

Zdroj: [1]

1.2.2 Jednoduchý graf

Jednoduchý graf je neorientovaný graf bez slučiek (self-edge) a viacnásobných hrán (multiedge) (Obr. 1.2) [1, 2]. V *neorientovanom grafe* hrany nemajú orientáciu, teda platí, že ak existuje hrana e_{ij} , existuje aj hrana e_{ji} [4]. *Slučka* je hrana e_{ii} , ktorá spája jeden a ten istý vrchol. *Viacnásobná hrana* vzniká, ak je medzi dvoma vrcholmi viac ako jedna hrana.



Obr. 1.2: Porovnanie jednoduchého grafu (naľavo) a grafu so slučkami a viacnásobnými hranami (napravo).

Zdroj: [1]

1.2.3 Ohodnotený graf

Ohodnotený (alebo aj vážený) graf má hrany alebo vrcholy ohodnotené číslom (váhou). Váha hrán je väčšinou reálne číslo vyjadrujúce silu prepojenia alebo vzdialenosť vrcholov, ktoré hrana spája [1]. Naopak hrany *nehodnoteného grafu* nemajú priradené žiadne váhy. Sú teda považované za rovnako dôležité alebo rovnako vzdialené od ostatných hrán.

1.2.4 Matica susednosti

Fundamentálnou matematickou reprezentáciou grafu je matica susednosti (Obr. 1.3). Matica susednosti A nehodnoteného grafu $G = (V, E)$ s n vrcholmi je definovaná ako $n \times n$ matica s prvkami A_{ij} , takými že [1]:

$$A_{ij} = \begin{cases} 1 & \text{ak hrana } e_{ij} \in E \\ 0 & \text{ak hrana } e_{ij} \notin E \end{cases} \quad (1.1)$$

V matici susednosti A jednoduchého grafu (Obr. 1.3) platí, že prvky na diagonále sa rovnajú nule ($A_{ii} = 0$), pretože graf neobsahuje slučky. Ďalej platí, že matica je symetrická ($A_{ij} = A_{ji}$), pretože ide o neorientovaný graf [1].

Hranovo ohodnotené grafy môžu byť taktiež reprezentované ako matica susednosti s prvkami rovným váham zodpovedajúcich hrán [1].

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Obr. 1.3: Matica susednosti jednoduchého neohodnoteného grafu.

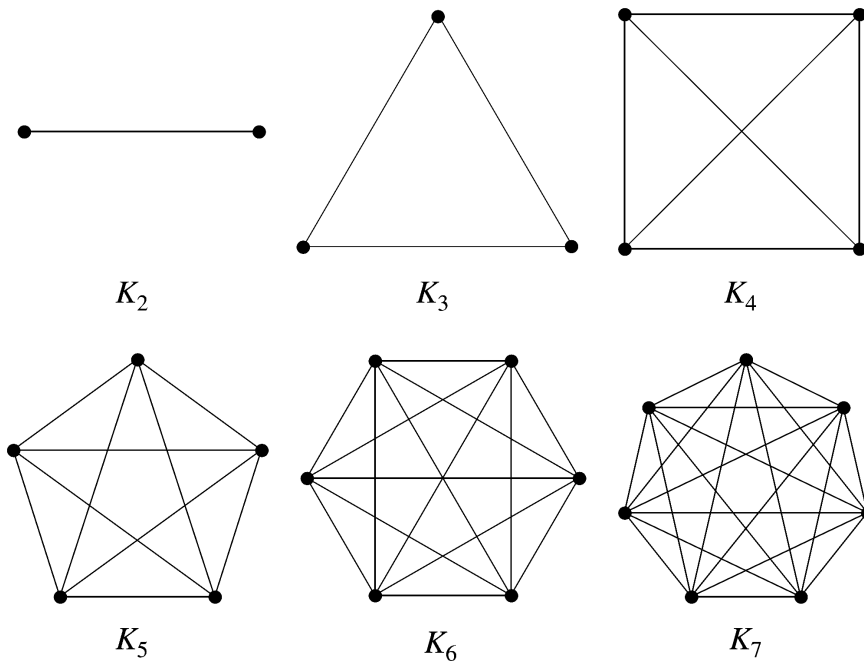
Zdroj: [1]

1.2.5 Kompletný graf

Kompletný graf je jednoduchý graf v ktorom platí, že všetky dvojice vrcholov sú navzájom susedné (Obr. 1.4). Hrana teda existuje medzi každým párom vrcholov. Kompletný graf s n vrcholmi označujeme K_n . Mohutnosť množiny hrán E kompletného grafu K_n , teda počet hrán takéhoto grafu vypočítame ako [2, 3, 5]:

$$|E| = \binom{n}{2} = \frac{n \cdot (n - 1)}{2} \quad (1.2)$$

Opakom (presnejšie grafovým doplnkom) kompletných grafov môžeme považovať takzvané prázdne grafy, ktoré neobsahujú žiadne hrany. Prázdny graf s n vrcholmi teda obsahuje len n izolovaných vrcholov a označujeme ho $\overline{K}_n$ [5]. *Izolovaný vrchol* je vrchol, ktorý nemá žiadne susedné vrcholy.



Obr. 1.4: Kompletne grafy s rôznym počtom vrcholov.

Zdroj: [5]

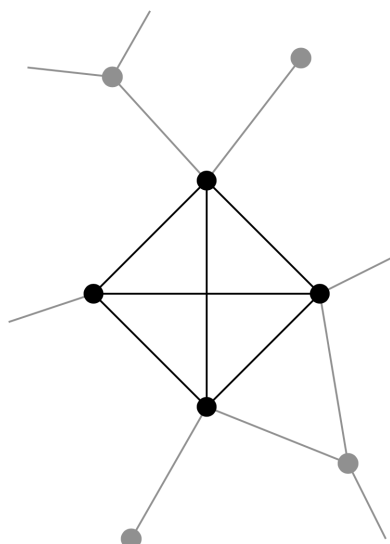
1.2.6 Podgraf

Jednoducho povedané, podgraf je časť grafu, ktorá vznikne z pôvodného grafu vymazaním niektorých jeho vrcholov alebo hrán. Ak však vymažeme vrchol, musíme zároveň vymazať všetky hrany vedúce do tohto vrcholu. Formálnejšie môžeme podgraf H grafu G zdefinovať tak, že platí [6]:

$$V(H) \subseteq V(G) \wedge E(H) \subseteq E(G) \quad (1.3)$$

1.2.7 Klika

Klika (clique) je množina vrcholov v neorientovanom grafe, v ktorej sú všetky vrcholy navzájom susedné (Obr. 1.5). Takáto množina teda spĺňa dve vlastnosti, je to podgraf neorientovaného grafu, a zároveň je to kompletný graf [7]. Kliky sa môžu v grafe prekrývať, čo znamená, že viaceré kliky môžu mať jeden alebo viac spoločných vrcholov [1].



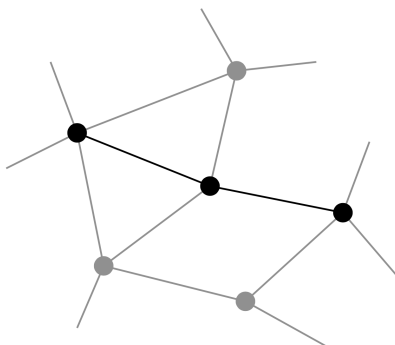
Obr. 1.5: Klika (čierna farba) v neorientovanom grafe, ktorá obsahuje 4 vrcholy.

Zdroj: [1]

1.2.8 Cesta

Na to, aby sme mohli zdefinovať cestu, musíme najprv zdefinovať sled. *Sled (walk)* je postupnosť $v_1e_1v_2e_2v_3e_3 \dots v_n$ kde v_i je vrchol pre $i \in \{1, 2, \dots, n\}$ a e_i je hrana pre $i \in \{1, 2, \dots, n-1\}$ [3]. Inými slovami, sled postupuje z vrcholu na susedný vrchol cez hranu, ktorá ich spája.

Cesta (path) z vrcholu u do vrcholu v je sled s koncovými vrcholmi u a v , v ktorom sa vrcholy ani hrany neopakujú (Obr. 1.6) [2, 3]. *Dĺžka cesty* je počet hrán v jej postupnosti [1, 2, 3].

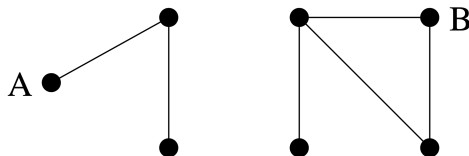


Obr. 1.6: Cesta (čierna farba) medzi dvoma vrcholmi, ktorá je zároveň aj ich najkratšia možná.

Zdroj: [1]

1.2.9 Komponent

Ak medzi každou dvojicou vrcholov grafu existuje cesta, takýto graf voláme *súvislý* (*connected*) [3]. Komponent je podmnožinou vrcholov grafu, v ktorej platí, že existuje cesta medzi každou dvojicou vrcholov a zároveň už nie je možné do podmnožiny pridať žiadny ďalší vrchol pre zachovanie tejto vlastnosti. Graf, ktorého všetky vrcholy patria do rovnakého jediného komponentu, je teda súvislý. Naopak graf s viac ako jedným komponentom je *nesúvislý* (Obr. 1.7) [1].



Obr. 1.7: Nesúvislý graf s dvoma komponentmi A a B.

Zdroj: [1]

1.2.10 Priemerná najkratšia dĺžka cesty

Uvažujme súvislý neorientovaný graf $G = (V, E)$ s n vrcholmi, v ktorom d_{uv} je *najkratšia cesta* (nazývaná aj *vzdialenosť*) medzi vrcholmi u a v (Obr. 1.6). Medzi každou dvojicou vrcholov v grafe G existuje najkratšia cesta. Dĺžku najdlhšej z týchto najkratších ciest definujeme ako *priemer grafu* G [1, 2].

Priemernú najkratšiu dĺžku cesty medzi vrcholom u a každým ďalším vrcholom v v grafe G definujeme ako [1]:

$$\ell_u = \frac{1}{n} \cdot \sum_{v \in V} d_{uv} \quad (1.4)$$

Priemernú najkratšiu dĺžku cesty medzi všetkými vrcholmi v grafe G , teda priemernú najkratšiu dĺžku cesty grafu G potom vyjadríme ako [1]:

$$\ell_G = \frac{1}{n} \cdot \sum_{v \in V} \ell_v = \frac{1}{n^2} \cdot \sum_{u, v \in V} d_{uv} \quad (1.5)$$

1.2.11 Stupeň vrcholu

Stupeň vrcholu je v teórii komplexných sietí jednou zo základných metrík pre meranie *centrálnosti (dôležitosti)* vrcholu v sieti [1]. Stupeň vrcholu $d(v)$ vyjadruje počet hrán pripojených k vrcholu [1, 2, 3]. Pomocou matice susednosti A (1.1) môžeme stupeň vrcholu $d(v)$ v grafe $G = (V, E)$ vyjadriť ako [1]:

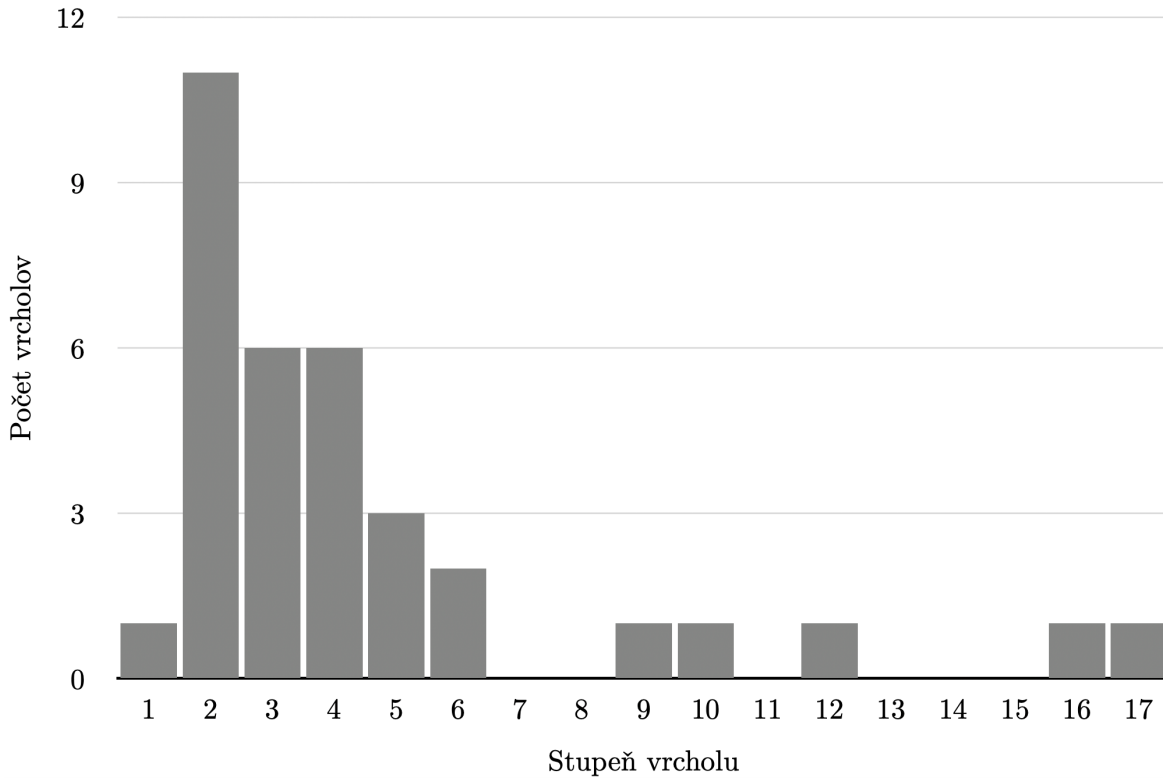
$$d(v) = \sum_{i \in V} A_{vi} \quad (1.6)$$

V jednoduchom neohodnotenom grafe platí, že stupeň vrcholu $d(v)$ je rovný mohutnosti množiny susedov vrcholu $N(v)$, teda:

$$d(v) = |N(v)| \quad (1.7)$$

1.2.12 Distribúcia stupňov

Tak ako stupeň (1.6) je základnou metrikou vrcholov, aj distribúcia stupňov je jednou z najzákladnejších a najdôležitejších charakteristík sietí. Distribúcia stupňov nám hovorí, s akou frekvenciou sa v sieti objavujú vrcholy s rôznymi stupňami [1]. Distribúcia stupňov sa najčastejšie vizualizuje pomocou stĺpcového grafu (Obr. 1.8).



Obr. 1.8: Vizualizácia distribúcie stupňov *Zachary's Karate Club* siete (Obr. 1.1).

Zdroj: vlastný

1.2.13 Klasterizačný koeficient

Klasterizačný koeficient vrcholu v nazývaný aj lokálny klasterizačný koeficient je podiel počtu hrán medzi susedmi vrcholu v a počtu všetkých možných hrán medzi susedmi vrcholu v [2]. Tento koeficient teda vyjadruje, s akou pravdepodobnosťou sú dvaja susedia vrcholu navzájom taktiež spojení hranou. V reálnom svete sa môžeme pýtať, s akou pravdepodobnosťou sú moji priatelia taktiež priatelia medzi sebou [4]. V neorientovanom grafe $G = (V, E)$ s n vrcholmi je klasterizačný koeficient vrcholu v daný ako [8]:

$$C_v = \frac{2 \cdot |\{e_{ij} : v_i, v_j \in N(v) \wedge e_{ij} \in E\}|}{|N(v)| \cdot (|N(v)| - 1)} \quad (1.8)$$

kde $N(v)$ je množina susedov vrcholu v .

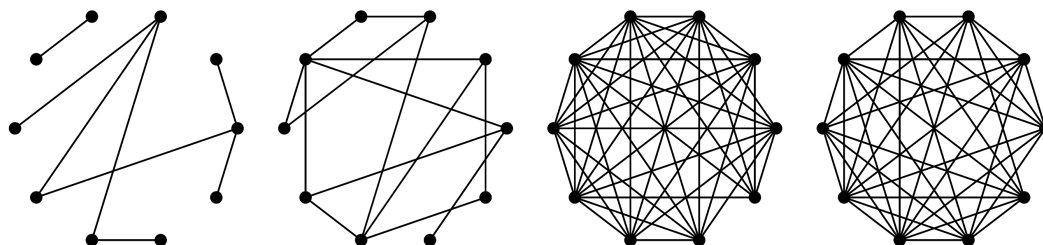
Globálny klasterizačný koeficient grafu G môže byť okrem iného definovaný aj ako priemerný klasterizačný koeficient všetkých jeho vrcholov [4, 9]:

$$C_G = \frac{1}{n} \cdot \sum_{v \in V} C_v \quad (1.9)$$

1.2.14 Náhodné grafy

Náhodný graf je graf, ktorého vlastnosti sú určené určitým náhodným spôsobom [5]. Najčastejšie sa zameriava na spôsob akým sú jednotlivé vrcholy prepojené hranami.

Za najznámejší model náhodných grafov sa pokladá *Erdős–Rényi model*, označovaný ako $G(n, M)$, kde n je počet vrcholov a M je počet hrán v grafe [10]. Graf začína s n izolovanými vrcholmi, pričom následne je každý jeden pár vrcholov spojený hranou s pravdepodobnosťou $p \in [0, 1]$ (Obr. 1.9). Graf môžeme teda označiť aj ako $G(n, p)$, kde n je počet vrcholov a p je pravdepodobnosť spojenia dvoch vrcholov hranou [11, 12].

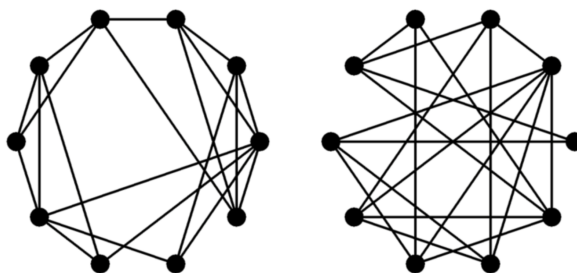


Obr. 1.9: Náhodné grafy s 10 vrcholmi $G(10, p)$ s pravdepodobnosťami p v intervale $[0, 1]$.

Zdroj: [5]

1.2.15 Sieť malého sveta

Efekt malého sveta bol objavený štúdiom sociálnych sietí v sociológii [8] a prvý krát bol popísaný vďaka slávnemu *Milgramovému experimentu* z roku 1967 v Nebraska, USA [4, 13]. Skupine ľudí boli rozoslané listy s rovnakým adresátom, ktorému bolo treba list doručiť. Ak človek adresáta poznal, poslal mu list priamo. Ak ho nepoznal, list poslal niekomu, o ktorom si myslel, že by adresáta mohol poznať. Každé zaslanie listu ďalšiemu človeku zvýšilo vzdialenosť cesty listu o jedna. Výsledná stredná hodnota vzdialenosti ciest bola rovná šiestim [2]. Tento výsledok je pôvodom známeho pojmu šesť stupňov odlúčenia (six degrees of separation), teda myšlienky, že medzi akýmkoľvek dvoma ľuďmi na svete existuje len približne šesť krokov [1, 4]. Milgramov experiment je presvedčivou demonštráciou efektu malého sveta, ktorý bol viac krát zrekonštruovaný a potvrdený aj skúmaním počítačových sociálnych sietí [8].



Obr. 1.10: Sieť malého sveta (naľavo) a náhodný graf (napravo) s rovnakým množstvom vrcholov (10) a hrán (v priemere 4 hrany na vrchol).

Zdroj: [14]

Z pohľadu teórie komplexných sietí, siete malého sveta majú vysoký klasterizačný koeficient (1.9), pomerne malý priemer a krátku priemernú najkratšiu dĺžku cesty (1.5) [1, 2, 4, 8]. Siete malého sveta majú na rozdiel od náhodných grafov (1.2.14) s rovnakým množstvom vrcholov a hrán jasne odlišnú topológiu (Obr. 1.10) [14]. Udáva sa, že ich priemerná najkratšia dĺžka cesty rastie úmerne logaritmu počtu ich vrcholov [1, 2]:

$$\ell_G \propto \log n \quad (1.10)$$

Ďalej sa udáva, že priemer siete sa vzhľadom na počet vrcholov správa podobne ako priemerná najkratšia dĺžka cesty [2]. Priemer je však vo všeobecnosti menej užitočnou metrikou, pretože meria vzdialenosť len dvoch konkrétnych vrcholov a môže byť podstatne ovplyvnený len jedným alebo malou skupinou vrcholov [1].

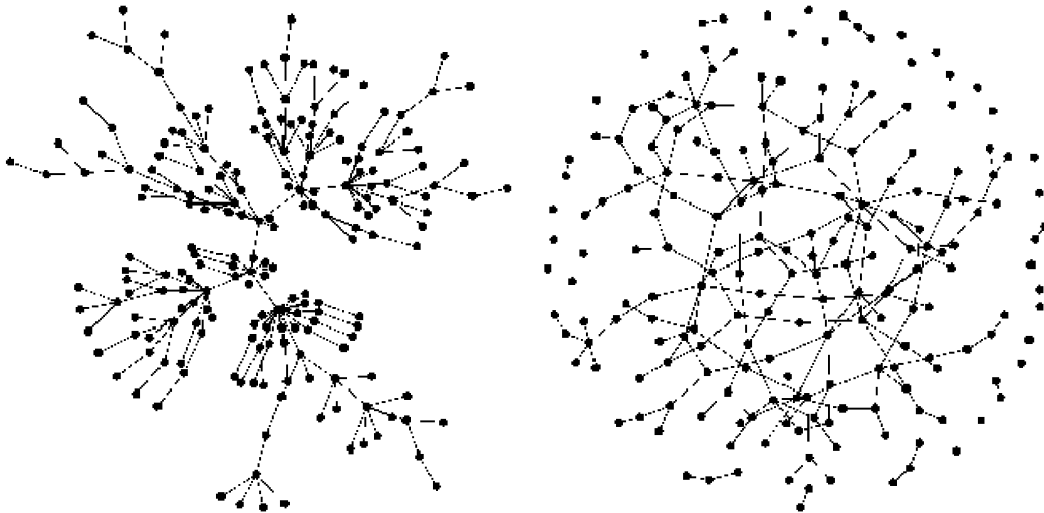
1.2.16 Bezškálové siete

Bezškálové siete nazývame siete, ktoré majú *mocninovú distribúciu stupňov* (*power-law distribution*) (Obr. 1.12) [1, 2, 4, 8]. Mocninová distribúcia stupňov sa riadi mocninovým zákonom, teda pravdepodobnosť že ľubovoľný vrchol má k susedov je približne [2, 4]:

$$p(k) \sim k^{-\gamma} \quad (1.11)$$

kde γ je typicky v intervale $2 \leq \gamma \leq 3$.

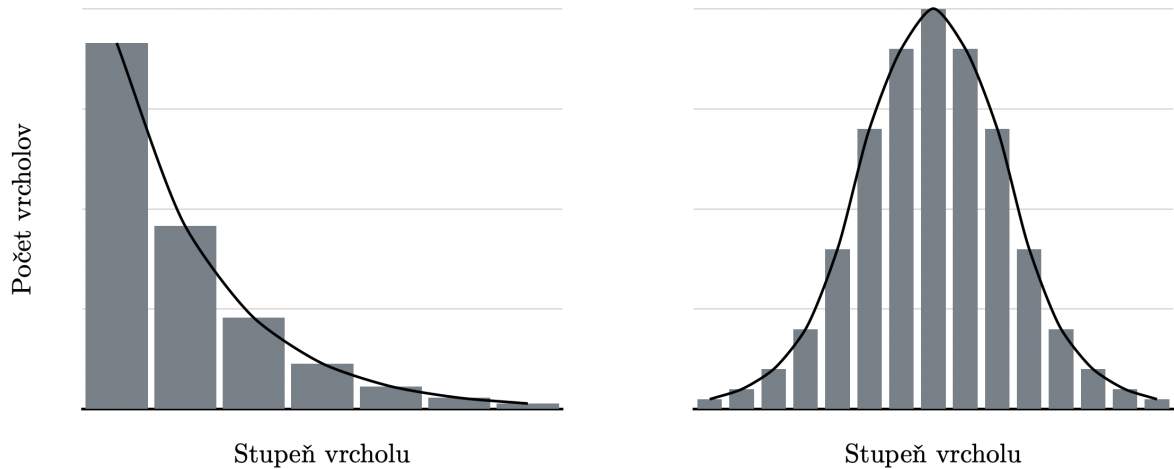
V bezškálových sieťach má väčšina vrcholov menej susedov, zatiaľ čo malé množstvo vrcholov má veľmi veľké množstvo susedov. Takéto vrcholy s veľkým množstvom susedov voláme *huby*, vďaka ktorým drží celá sieť pohromade (Obr. 1.11) [4].



Obr. 1.11: Porovnanie bezškálovej siete s hubmi (naľavo) a náhodného grafu (napravo).

Zdroj: [8]

Vďaka týmto vlastnostiam sú bezškálové siete odolné voči náhodným poruchám (náhodne odstránenie vrcholu či hrany). Naopak sú však veľmi zraniteľné voči koordinovaným útokom [4].



Obr. 1.12: Mocninová distribúcia stupňov (naľavo) a binomická distribúcia stupňov (napravo).

Zdroj: vlastný

Náhodne grafy (1.2.14) majú na rozdiel od bezškálových sietí *binomickú distribúciu stupňov* (Obr. 1.12), teda pravdepodobnosť že ľubovoľný vrchol v grafe s n vrcholmi má k susedov je [1, 8]:

$$p(k) \approx \binom{n-1}{k} \cdot p^k \cdot (1-p)^{n-1-k} \quad (1.12)$$

kde p je pravdepodobnosť prítomnosti hrany.

1.2.17 Komplexné siete

Pre komplexné siete, často nazývané aj prirodzené či reálne siete neexistuje formálna definícia [2]. Komplexné siete prirodzene vznikajú a môžeme ich pozorovať všade vo svete okolo nás. Medzi najznámejšie príklady patria sociálne siete priateľstiev, informačné siete spoluvýskytu slov, technologická sieť internetu, biologické siete interakcií proteínov a mnoho ďalších [1]. Takéto siete alebo netriviálne systémy boli najprv považované za náhodne. Neskôr sa však zistilo, že rozličné typy sietí z reálneho sveta majú podobné vlastnosti, vďaka čomu sú získané znalosti a navrhnuté algoritmy pre konkrétny typ komplexnej siete pravdepodobne použiteľné aj pre iné komplexné siete [8].

Medzi typické charakteristiky komplexných sietí patrí ich štruktúra, ktorá nie je presne usporiadaná, no nie je ani náhodná. Komplexné siete ďalej bývajú veľké, dynamicky sa menia a spĺňajú definície sietí malých svetov a bezškálových sietí [2, 4, 8]. Vďaka týmto vlastnostiam fungujú algoritmy na dolovanie alebo spracovanie

údajov v sieťach a kategorizácia na základe topologickej štruktúry sietí, ako napríklad PageRank [8].

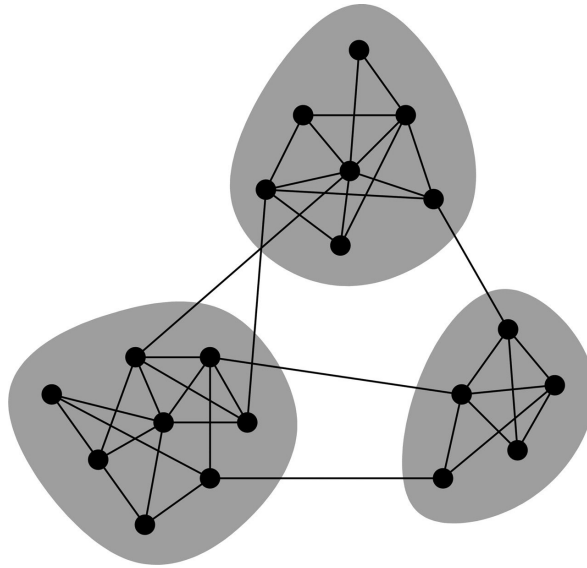
Známy príklad malej (34 vrcholov a 78 hrán) reálnej sociálnej siete je *Karate Club Waynea Zacharyho* (Obr. 1.1) zo sociologickej literatúry, ktorá bola popísaná v článku o konfliktoch a štiepení v malých sociálnych skupinách [15]. Táto sieť predstavuje priateľstvá medzi členmi klubu karate v severoamerickej univerzite [1]. Výskumníci a vedci často používajú túto sieť ako *benchmark* na testovanie a vývoj algoritmov týkajúcich sa analýzy komplexných sietí ako napríklad detekcia komunít, ktorú popíšeme v nasledujúcej podsekcii.

1.2.18 Detekcia komunít

Detekcia komunít je jedna z najaktívnejších oblastí súčasného výskumu komplexných sietí [1]. Komunity sú *zhluky* (*clusters*) alebo skupiny vrcholov v sieti, ktoré sú silne prepojené medzi sebou a slabšie prepojené so zvyškom siete. Najčastejšie používané definície sú založené na počte hrán v rámci komunity a počte hrán vedúcich mimo komunitu (Obr. 1.13) [8].

Schopnosť detekovať komunity v sieťach môže byť veľmi užitočným nástrojom na odhaľovanie ich štruktúry a organizácie. Ako príklad môžeme uviesť sieť spoluautorstiev na univerzitnej katedre, kde vrcholy tvoria vedci a hrany spájajú dvojice vedcov, ktorí sú spoluautormi vedeckých prác. Komunity v tomto prípade zodpovedajú skupinám vedcov, ktorí spolu úzko spolupracovali. Takáto sieť má teda komunitnú štruktúru, ktorá odráža výskumné skupiny a rozdelenie záujmov v rámci katedry [1].

Ďalším známym príkladom použitia algoritmov na detekciu komunít je odhaľovanie teroristických skupín na sociálnych sieťach sledovaním ich aktivít a interakcií [16].



Obr. 1.13: Komunity (šedé oblasti) v malom grafe.

Zdroj: [17]

1.2.19 Centrálnosť blízkosti

Podobne ako stupeň vrcholu (1.6), aj centrálnosť blízkosti (closeness centrality) je metrikou pre meranie centrálnosti alebo dôležitosti vrcholu v sieti. Táto metrika je však v porovnaní so stupňom špecifickejšia a oveľa zaujímavejšia. Vyjadruje to, ako blízko sa vrchol nachádza od všetkých ostatných vrcholov v sieti. Napríklad v sociálnej sieti priateľstiev sa názory osoby s vysokou centrálnosťou blízkosti šíria oveľa rýchlejšie ako názory iných osôb [1].

Z matematického hľadiska môžeme definovať centrálnosť blízkosti pomerne jednoducho, a to pomocou priemernej najkratšej dĺžky cesty ℓ_v (1.4) medzi vrcholom v a každým ďalším vrcholom v súvislom a neohodnotenom grafe. Keďže pre najcentrálnejšie vrcholy chceme dostávať najvyššie hodnoty, centrálnosť blízkosti vrcholu je definovaná ako prevrátená hodnota [1, 2]:

$$C_v = \frac{1}{\ell_v} \quad (1.13)$$

Kapitola 2

Vybrané postupy pre zber dát, analýzu a vizualizáciu

Vybrané postupy.

2.1 Zber dát

2.2 Extrakcia kľúčových slov

2.3 Analýza sentimentu

2.4 Modelovanie komplexných sietí

2.5 Analýza a modelovanie tém

2.6 Vizualizácia

2.7 Limity navrhnutých postupov

Kapitola 3

Softvérový nástroj

3.1 Použité technológie

3.2 Implementácia

Kapitola 4

Výskum výstupov

V nasledujúcej kapitole prezentujeme a skúmame výstupy zo softvérového nástroja pre rôzne zaujímavé vyhľadávacie dotazy, ktoré sa týkajú viacerých aktuálnych oblastí a tém vo svete a našej spoločnosti.

Záver

Literatúra

- [1] Mark Newman. *Networks*. Oxford University Press, 2018.
- [2] Mária Markošová a Peter Náther. *Prednáškové prezentácie z predmetu komplexné siete (2-AIN-154)*. Univerzita Komenského – Fakulta matematiky, fyziky a informatiky, Bratislava.
- [3] Edita Máčajová. *Prednáškový materiál z teórie grafov (2-INF-174)*. Univerzita Komenského – Fakulta matematiky, fyziky a informatiky, Bratislava, 2021.
- [4] Kristína Dvorská. *Analýza vytvárania sociálnych sietí a ich vlastností (Bakalárska práca)*. Univerzita Komenského – Fakulta matematiky, fyziky a informatiky, Bratislava, 2021.
- [5] Eric Weisstein. *Wolfram MathWorld – The web’s most extensive mathematics resource*, 2023. <https://mathworld.wolfram.com>.
- [6] Lukáš Jirovský. *Teorie grafů ve výuce na střední škole (Diplomová Práce)*. Univerzita Karlova v Praze – Matematicko-fyzikální fakulta, 2010.
- [7] Beáta Stehlíková. *Prednáškový materiál z predmetu analýza sociálnych sietí (2-EFM-155)*. Univerzita Komenského – Fakulta matematiky, fyziky a informatiky, Bratislava, 2020.
- [8] Marek Ciglan a Michal Laclavík. *Detekcia komunit v komplexných sieťach*. Ústav informatiky SAV, 2013.
- [9] NetworkX Developers. *NetworkX dokumentácia*, 2022. <https://networkx.org/documentation/stable>.
- [10] Béla Bollobás. *Random Graphs*. Academic Press Inc., 1985.
- [11] Kristína Dvorská. *Analýza sociálnych sietí (Diplomová práca)*. Univerzita Komenského – Fakulta matematiky, fyziky a informatiky, Bratislava, 2023.
- [12] Ernesto Estrada. *The structure of complex networks: theory and applications*. Oxford University Press, 2012.

- [13] Jeffrey Travers a Stanley Milgram. *An Experimental Study of the Small World Problem*. Sociometry, 1969.
- [14] Nichol Castro a Cynthia Siew. *Contributions of modern network science to the cognitive sciences: Revisiting research spirals of representation and process*. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2020.
- [15] Wayne Zachary. *An Information Flow Model for Conflict and Fission in Small Groups*. Journal of Anthropological Research, 1977.
- [16] Ivan Despot. *Memgraph's guide for NetworkX library*, 2022. <https://networkx.guide>.
- [17] Mark Newman. *Modularity and community structure in networks*. Proceedings of the National Academy of Sciences, 2006.