FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Zadání bakalářské práce

Název:	Analýza míry skupinové centrality reálných sociálních sítí
Student:	Matyáš Turek
Vedoucí:	Ing. Šimon Schierreich
Studijní program:	Informatika
Obor / specializace:	Znalostní inženýrství
Katedra:	Katedra aplikované matematiky
Platnost zadání:	do konce letního semestru 2023/2024

Pokyny pro vypracování

1. Nastudujte různé míry skupinové centrality sociálních sítí, které jsou studovány v literatuře, např. [1,2,3].
2. Implementujte různé míry skupinové centrality z bodu 1.
3. Naměřte hodnoty těchto měr pro alespoň pět datasetů reálných sociálních sítí [4].
4. Diskutujte výsledky naměřené v bodě 3.

[1] M. G. Everett & S. P. Borgatti (1999) The centrality of groups and classes, The Journal of Mathematical Sociology, 23:3, 181-201, DOI: 10.1080/0022250X.1999.9990219
[2] E. D. Kolaczyk, D. B. Chua & M. Barthélemy (2009) Group betweenness and co-betweenness: Inter-related notions of coalition centrality, Social Networks, 31:3, 190-203, DOI: 10.1016/j.socnet.2009.02.003.
[3] A. Veremyev, O. A. Prokopyev & E. L. Pasiliao (2017) Finding groups with maximum betweenness centrality, Optimization Methods and Software, 32:2, 369-399, DOI: 10.1080/10556788.2016.1167892
[4] J. Leskovec & A. Krevl (2014) SNAP Datasets: Stanford Large Network Dataset Collection [online], URL: https://snap.stanford.edu/data/index.html#socnets

Elektronicky schválil/a Ing. Magda Friedjungová, Ph.D. dne 21. listopadu 2022 v Praze.

FAKULTA
INFORMAČNÍCH
TECHNOLOGIÍ
ČVUT V PRAZE

Bakalářská práce

Analýza míry skupinové centrality reálných sociálních sítí

Matyáš Turek

Katedra aplikované matematiky
Vedoucí práce: Ing. Šimon Schierreich

10. května 2023

Poděkování

Mé poděkování patří Ing. Šimon Schierreichovi za odborné vedení práce, cenné rady a ochotnou pomoc, díky které jsem dokázal tuto práci vypracovat.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principu při přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisu. V souladu s ust. § 2373 odst. 2 zákona č. 89/2012 Sb., občanský zákoník, ve znění pozdějších předpisu, tímto uděluji nevýhradní oprávnění (licenci) k užití této mojí práce, a to včetně všech počítačových programu, jež jsou její součástí či přílohou a veškeré jejich dokumentace (dále souhrnně jen „Dílo"), a to všem osobám, které si přejí Dílo užít. Tyto osoby jsou oprávněny Dílo užít jakýmkoli zpusobem, který nesnižuje hodnotu Díla a za jakýmkoli účelem (včetně užití k výdělečným účelum). Toto oprávnění je časově, teritoriálně i množstevně neomezené. Každá osoba, která využije výše uvedenou licenci, se však zavazuje udělit ke každému dílu, které vznikne (byť jen zčásti) na základě Díla, úpravou Díla, spojením Díla s jiným dílem, zařazením Díla do díla souborného či zpracováním Díla (včetně překladu) licenci alespoň ve výše uvedeném rozsahu a zároveň zpřístupnit zdrojový kód takového díla alespoň srovnatelným zpusobem a ve srovnatelném rozsahu, jako je zpřístupněn zdrojový kód Díla.

V Praze dne 10. května 2023 . . . . . .

České vysoké učení technické v Praze
Fakulta informačních technologií

Odkaz na tuto práci

Turek, Matyáš. *Analýza míry skupinové centrality reálných sociálních sítí.* Bakalářská práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2023.

Abstrakt

Tato bakalářská práce se zabývá detailním vysvětlením skupinových měr centralit, což je jedna z technik analýzi grafů sítí. Jednotlivé míry jsou zde popsány z hlediska využití a také vysvětleny pricipy algoritmů na zjištění těchto měr.

Na reálných datasetech sociálních sítí jsou naměřeny jednotlivé míry a k těmto výsledkům je vedena diskuze.

Klíčová slova skupinová míra centrality, sociální sítě, měření centrality, teorie grafů, reálné datasety, algoritmy grafové teorie

Abstract

This bachelor thesis deals with a detailed explanation of group centrality measures, which is one of the techniques for analyzing graph networks. The various measures are described in terms of their applications and the principles of the algorithms for finding these measures are also explained.

The individual measures are measured on real social network datasets and a discussion is given on these results.

Keywords group centrality, social networking sites, measuring centralities, graph theory, real datasets, graph theory algorithms

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Sociální síť, společenská síť nebo komunitní síť je služba na Internetu, která registrovaným členům umožňuje si vytvářet osobní (či firemní) veřejný či částečně veřejný profil, komunikovat spolu, sdílet informace, fotografie, videa, provozovat chat a další aktivity. Pojmenování pochází ze sociologického pojmu sociální síť – skupina lidí, která spolu udržuje komunikaci různými prostředky [1].

Sociální sítě mohou být také různá diskusní fóra, bazary, služby, kde uživatelé přidávají různá videa a jiné, podle toho se dělí na profilově orientované jako například Facebook a Google+ a obsahově orientované jako je Youtube a Instagram. Celkově sociální sítě v této době využívá něco málo přes 3,4 miliardy uživatelů a tak se staly nedílnou součástí života.

Sociální sítě jdou velmi jednoduše znázornit pomocí grafové teorie, jednotliví uživatelé jsou vrcholy grafu a hrany tohoto grafu mohou představovat buď různé vztahy v rámci sociální sítě jako například přátelství, nebo také interakce jako například okomentování příspěvků jiného autora atd.

Pro řadu profesí je důležité mít představu o tom, jak jsou jednotliví aktéři v rámci sociálních sítí vlivní. Ač je tento parametr velmi diskutabilní, tak existuje řada nástrojů, které se jej snaží kvantifikovat. Často jejich konkrétní hodnoty neodpovídají zcela přesně skutečné situaci, ale přinejmenším dokáží zachytit trendy relativně dobře [2]. Jedním z takových možných měření je měření centrality.

Centralita je pojem užívaný v oblasti teorie grafů a síťové analýze, popisující provázanost jednotlivých uzlových vrcholů daného systému. Tato koncepce se využívá pro analýzu vztahů v rámci sítě (například sociální), k vyjádření důležitosti uzlového vrcholu jako takového (například člověka, skupiny atp.), a především pak k popisu jeho umístění v soustavě. Díky této veličině se tedy dá určit míra propojení tohoto vrcholu se zbytkem systému a jeho celkové začlenění. [3]. Pokud tedy chceme hledat důležitou osobu, stránku nebo skupinu v rámci sítě, míry centrality nám pomohou odpovědět na tento problém.

Obecně mít nějaký nejcentrálnější prvek v rámci sítě je velmi důležité. Předpokládáme například, že ústředí firmy bude to nejcentrálnější místo v rámci poboček dané firmy, nebo například, že v rámci dané politické strany bude její předseda nejcentrálnější osoba. U sociálních sítí ale většinou nemáme přesně dané pozice, abychom mohli hned prohlásit, že daný jedinec bude tím nejdůležitějším prvkem. Centrální pozice jsou vždy ztotožňovány se správným vedením, dobrou popularitu nebo vynikající pověst v síti [4].

Důležitost v rámci sítě ale může znamenat mnohé a dle našeho konkrétního požadavku nám může vyhovovat něco jiného. Proto také existují různé míry centralit. Nejobecněji můžeme rozdělit míry centrality jako vnímání toku informací skrz síť nebo jako pozici vrcholu v rámci soudržnosti sítě. V našem případě budeme uvažovat pouze důležitost v rámci soudržnosti sítě. Ale i tu můžeme rozdělit na několik dalších, v závislosti na naších požadavcích.

V roce 1999 přišli Martin G. Everett a Stephen P. Borgatti s rozšířením klasicky používaných měr centralit s rozšířením na skupinové míry centralit [5]. Původně síťoví analytici používali pouze míry centralit, které se vztahovaly k jednotlivých aktérům v rámci sítě, ale v mnoha případech a otázkách, které byly kladeny, nebylo možné použít klasickou míru centrality a tak toto rozšíření bylo velmi důležité.

Skupinové míry centrality nyní mohou odpovědět na dotazy jako například kdo je nejcentrálnější skupinou ve firmě, jestli to jsou například obchodní zástupci, či marketingové oddělení. Nebo také lze tímto rozšířením řešit i inverzní problém a to když by manažer chtěl sestavit tým z konkrétního počtu lidí, tak aby se mu podařilo sestavit tým, který bude ten nejcentrálnější.

Cíl práce

Prvně si v této práci představíme nutnou teorii k pochopení a významu skupinových měr centralit, to znamená úvod do teorie grafů a algoritmy, které budeme k centralitám potřebovat. Dále také v práci představíme a implementujeme algoritmy na výpočet jednotlivých skupinových měr centralit. z uvedeného zdroje[1] vybereme 5 datasetů sociálních sítí, na kterých naměříme zmíněné skupinové míry centrality. Naměřené výsledky poté prodisktujeme.

[1] https://snap.stanford.edu/data/index.html#socnets

Kapitola 1

Definice a pojmy

1.1 Teorie grafů

V této podkapitole budou zavedeny základní pojmy a definice, které jsou nutné pro formální zavedení měr centralit. Pro notaci grafové teorie vycházíme z monografií Diestel [6] a Matoušek a Nešetřil [7].

Graf je dvojice množin $G = (V, E)$ taková, že $E \subseteq [V]^2$, tedy prvky E jsou neuspořádané dvouprvkové podmnožiny V. Prvky V nazýváme *vrcholy* grafu G a prvky E nazýváme hrany, kdy jedna hrana spojuje vždy 2 vrcholy. Množinu všech vrcholů grafu G označujeme $V(G)$ a množinu jeho hran pak $E(G)$. Nechť $e = \{u, v\}$ je hrana grafu G, pak vrchol u a v jsou *koncové vrcholy* hrany e. Dále také u je *sousedem* v a vrcholy *u, v* jsou *incidentní* s hranou e.

Grafy se dělí na konečné, nekonečné a spočetné, ale v této práci se budeme zabývat pouze konečnými grafy, to jsou grafy, které mají konečnou množinu vrcholů a hran. Graf může být také *orientovaný* a pro ten platí, že E je množina orientovaných hran neboli prvky množiny E jsou uspořádané dvouprvkové podmnožiny V. V naši práci budeme používat pouze neorientované grafy.

Počtu vrcholů v grafu nazýváme *řád grafu* a můžeme jej zapisovat $|G|$, běžnější je ale notace $|G| = n$ a $|E| = m$. Rozlišujeme také prázdný graf, pro který platí $G = \{\emptyset, \emptyset\}$. Dále také máme triviální graf, pro který platí, že jeho řád je 1.

Graf může být také *úplný*, to znamená, že v něm jsou každé dva vrcholy spojené hranou. Takový graf se potom označuje K^n, kde n je počet vrcholů. Z toho také plyne, že $m = \binom{n}{2}$.

Je-li U libovolná množina vrcholů grafu G píšeme $G - U$ pro $G[V \setminus U]$, takto označujeme množinový rozdíl, tj. množinu všech prvků, které leží v G, ale nikoli v U [8]. $G-U$ tedy získáme tedy tak, že odstraníme všechny vrcholu v $U \cap V$ a jejich incidentní hrany. Pokud má $U = \{v\}$, pak můžeme také psát $G - \{v\}$.

1.1.1 Sousedství

Nechť $G = (V, E)$ je neprázdný graf. Množina sousedů vrcholu v v G je taková množina, pro kterou platí $N(v) = \{u \mid \{u, v\} \in E\}$ zapisujeme $N_G(v)$, pokud je G jasný z kontextu, můžeme také označovat pouze $N(v)$ a nazýváme to jako *otevřené okolí* vrcholu v. Pro libovolnou podmnožinu vrcholů U sousedy v $V \setminus U$ označujeme jako $N(U)$.

Stupeň vrcholu v je velikost jeho množiny sousedů, toto čislo je rovno počtu incidentních hran tohoto vrcholu. Stupeň vrcholu zapisujeme jako $\deg_G(v)$, opět pokud známe kontext grafu tak můžeme zkrátit na $\deg(V)$. Průměrný stupeň vrcholu v grafu pak vypočítáme jako

$$\deg(G) := \frac{1}{|V|} \sum_{v \in V} \deg(v).$$

Pokud máme orientovaný graf, tak u vrcholu můžeme rozlišovat na vstupní stupeň vrcholu, což je počet orientovaných hran vedoucích do tohoto vrcholu, a výstupní stupeň vrcholu, což je počet orientovaných hran, které z tohoto vrcholu vedou [9].

Vrchol stupně 0 nazýváme *izolovaný*. Pokud všechny vrcholu grafu G mají stejný stupeň $r, r \in \mathbb{N}^+$, pak takový graf nazýváme r-regulární.

1.1.2 Cesta

Sled délky l v grafu G je sekvence $v_0, e_1, v_1, e_2, \ldots, e_l, v_l$ taková, že $e_i = \{v_{i-1}, v_i\}$ a $e_i \in E(G)$ pro $i \in \{1, \ldots, l\}$. Cesta v G je sled, ve kterém se neopakují vrcholy, tím pádem se v něm nemohou opakovat ani hrany. *Délka* cesty $u - v$ je rovna počtu hran v této cestě. [10]

Vzdálenost vrcholů u, v, značme $\mathrm{dist}(u, v)$, je délka nejkratší $u - v$ cesty v G. Pokud taková cesta neexistuje, pak $\mathrm{dist}(u, v) := \infty$. Nejvetší vzdálenost dvou vrcholů v G nazýváme průměr grafu.

1.1.3 Podgraf

Graf H je *podgrafem* grafu G, když $V(H) \subseteq V(G)$ a $E(H) \subseteq E(G)$. Podgraf H nazveme *indukovaný*, pokud $V(H) \subseteq V(G)$ a $E(H) = E(G) \cap \binom{V(H)}{2}$. Významný podgraf je také *klika*, což je podgraf G, který je úplný.

1.1.4 Souvislost

Graf G nazveme *souvislý*, pokud je neprázdný a pro všechny dvojice vrcholů tohoto grafu existuje cesta v G, jinak graf nazýváme *nesouvislý*. Indukovaný podgraf H grafu G nazveme souvislou *komponentou*, pokud je souvislý a neexistuje žádný souvislý podgraf H takový, že $H \neq H$ a zároveň $H \in H$ [6].

1.2 Asymptotická složitost

Pro porovnání efektivnosti algoritmu dle [11] existují hlavní 2 porovnání a to

- *Časová složitost* - doba výpočtu podle daného algoritmu a pro daný objem dat.
- *Paměťová složitost* - velikost paměti využívané při výpočtu.

Skutečnou složitost výpočtu není možné v obecném případě přesně spočítat, protože závisí na implementaci algoritmu a konkrétním počítači, na kterém se algoritmus provádí. Abychom alespoň něco mohli spočítat, začaly se používat odhady složitosti. Tyto odhady popisují, jak rychle se zvyšuje složitost vzhledem k rostoucím vstupům, ale nedávají konkrétní funkční hodnotu.

Asymptotická složitost je způsob dělení algoritmů podle operační náročnost algoritmu. Asymptotická složitost algoritmu vypovídá o tom, jakým způsobem se bude chovat algoritmus v závislosti na změně rozsahu vstupních dat [12]. Zapisujeme dle Landauovy notace.

Máme-li funkce $f(n)$ a $g(n)$, pak řekneme, že $f(n)$ je nejvýše řádu $g(n)$, zapisujeme $f(n) = \mathcal{O}(g(n))$, jestliže

$$\exists c \in \mathbb{R}^+ \quad \exists n_0 \in \mathbb{N}^+ \quad \forall n \geq n_0 : \quad f(n) \leq c \cdot g(n).$$

Máme-li funkce $f(n)$ a $g(n)$, pak řekneme, že $f(n)$ je nejméně řádu $g(n)$, zapisujeme $f(n) = \omega(g(n))$, jestliže

$$\exists c \in \mathbb{R}^+ \quad \exists n_0 \in \mathbb{N}^+ \quad \forall n \geq n_0 : \quad c \cdot g(n) \leq f(n).$$

Máme-li funkce $f(n)$ a $g(n)$, pak řekneme, že $f(n)$ je téhož řádu jako $g(n)$, zapisujeme $f(n) = \theta(g(n))$, jestliže

$$\exists c_1, c_2 \in \mathbb{R}^+ \quad \exists n_0 \in \mathbb{N}^+ \quad \forall n \geq n_0 : \quad c_1 \cdot g(n) \leq f(n) \leq c_2 \cdot g(n).$$

1.3 Normalizace

Normalizace dat je běžná lineární transformace, která hodnotu daného čísla převede do intervalu $[0, 1]$ podle následujícího vzorce

$$x = \frac{x - min_x}{max_x - min_x}.$$

Normalizace hodnoty nám může o hodnotě říct daleko více, než konkrétní číslo. Jelikož někdy neznáme kontext dané hodnoty, tak díky normalizaci můžeme odhadnout jestli se hodnota blíží možnému maximu - normalizovaná hodnota se blíží 1, nebo možnému minimu - normalizovaná hodnota se blíží k 0.

1.4 Algoritmus procházení do šířky

Algoritmus procházení do šířky, anglicky známý jako breadth first search algoritmus nebo zkráceně BFS, je záplavový algoritmus procházení grafu od počátečního, námi zvoleného, vrcholu a postupně prochází všechny sousední vrcholy po vrstvách, které jsou určené vzdáleností od počátečního vrcholu, dokud není splněná ukončovací podmínka, nebo algoritmus neprošel všechny dosažitelné vrcholy.

Pro vysvětlení algoritmu budeme potřebovat datovou strukturu fronta. Jedná se o datovou strukuturu typu FIFO, z anglického first-in first-out, přeloženo do čestiny jako první dovnitř, první ven. Tato datová struktura podporuje hlavně 2 operace a to metodu push, kdy přidá prvek na konec fronty a metodu pop, ta zase vyjme a vymaže první prvek z této fronty. Pseudokod algoritmu je zobrazen v ukázce BFS Algoritmus 1.

Algoritmus BFS budeme v této práci využívat hlavně ke zjištění vzdálenosti mezi 2 vrcholy. Toho BFS dosáhne nalezením cesty s nejkratší vzdáleností. Toto je pseudokod algoritmu vyhledávání do šířky.

BFS Algoritmus

```
procedure BFS( graf G, vrchol s)
    for každý vrchol v ∈ V(G) do:
        stav(v) := nenalezený
        dist(v) := předek(v) := undef
    end for
    Q := fronta obsahující jediný vrchol s
    stav(s) := otevřený
    dist(s) = 0
    while fronta Q je neprázdná: do
        vyjmi z Q první vrchol v
        for každý soused w vrcholu v do
            if stav(w) = nenalezený: then
                stav(w) := otevřený
                dist(w) := dist(v) + 1
                předek(w) := v
                přidej w na konec fronty Q
            end if
        end for
        stav(v) := uzavřený
    end while
end procedure
```

1.4.1 Vzorkování sociálních sítí

Jelikož datasety sociálních sítí dosahují obrovského množství dat a nemáme k dispozici potřebnou výpočetní techniku, bylo nutné si z datasetů vyvzorkovat data tak, aby co možná nejpřesněji odpovídala celému datasetu. K tomuto účelu existují mnohé přístupy, naším zvoleným bude vzorkování náhodnou procházkou.

Náhodná procházka grafem je proces, který začíná v libovolném vrcholu a v každém časovém kroku se přesune do jiného vrcholu. Vrchol, do kterého se procházka přesouvá, je vybrán náhodně ze sousedů současného vrcholu [13].

Technika vzorkování sociálních sítí je prozkoumaná metoda a je ukázáno že statisticky dokáže poměrně dobře zachovat vlastnosti grafu [14] [15]. Bylo také vyzkoušeno na našich testovacích datasetech, že vzorkování náhodnou procházkou velmi přesně zachovává vlastnosti skupinových měr centralit.

Kapitola 2

Míry centrality

Názvy jednotlivých centralit budeme označovat v jejich anglickém originále, jelikož v češtině některé z nich ztrácí svou přesnost. Nejznámější a nejpoužívanější míry centrality pro zjištění důležitosti vrcholů v rámci sítě, z nichž budou první 3 rozebrány později. Budeme se zaměřovat pouze na první 3, jelikož mají své rozšíření také na skupinovou míru centrality, kterými se tato práce zabývá. Obecně nejpoužívanější jsou [16]:

- Degree centrality
- Closeness centrality
- Betweenness centrality
- Eigenvector centrality
- PageRank centrality

2.1 Skupinové míry centralit

Kromě použití na apriorně určených skupinách, můžeme skupinové míry centralit použít také na skupiny jednotlivců získaných pomocí technik soudržných podskupin, jako jsou například kliky v grafu. Mohli bychom tedy první kliky identifikovat a poté je ohodnotit a nalézt tak tu nejcentrálnější. Ovšem najít kliky zadané velikosti v grafu je tzv. NP-těžký problém [17]. NP-těžký je takový problém, na který lze polynomiální redukcí převést jakýkoliv problém ze třídy NP. NP-těžký tedy znamená "alespoň tak těžký jako jakýkoli problém ve třídě NP", i když ve skutečnosti může být těžší [18]. Obecně se domníváme, že pro NP-těžké problémy neexistují algoritmy, které by jej řešili v polynomiálním čase. Další takové problémy z grafové teorie jsou například problém obarvitelnosti grafu [19] nebo problém hledání nejdelší cesty v grafu [20].

2.1.1 Degree centrality

Degree centrality je konceptuálně nejjednodušší a také celkově byla první navrženou mírou centrality. U této míry zkoumáme otevřené okolí vrcholu.

Pro klasickou degree centrality na individuálních vrcholech je míra daná velikostí jeho množiny sousedů, neboli jeho stupněm. Degree centrality pro individuální vrchol v tedy můžeme získat jako

$$C_d(v) = d_G(v).$$

Pokud bychom chtěli tuto míru normalizovat, využijeme následující vzorec.

$$C_d(v) = \frac{d_G(v)}{|G| - 1}$$

U orientovaných grafů lze také rozlišovat na *inlink* a *outlink degree centrality*. Rozdělení na inlink a outlink degree centrality se používá jako jedno z kritérií u PageRanku na webu. Stránky s vysokou mírou inlink degree centrality jsou označovány jako autority, mohou to být stránky, na které se často odkazuje, nebo se z nich často cituje atd., a stránky s vysokou mírou outlink degree centrality se označují jako huby, neboli rozbočovače [21].

Rozšíření na skupinovou degree centrality je zde velice prosté, zavedeme tento počet jako počet vrcholů z jiné skupiny, které jsou sousední s alespoň jedním vrcholem, který k dané skupině patří. Pokud máme tedy podmnožinu vrcholů grafu G, jako skupinu S, tak skupinovou degree centrality získáme jako

$$C_{gd}(S) = \{v : (u, v) \in E \wedge u \in S \wedge v \notin S.\}$$

Pokud bychom opět chtěli normalizovat tuto hodnotu, tak za předpokladu, že $C_{gd}(S)$ je již naše vypočítaná skupinová degree centrality, tak její normalizovanou hodnotu získáme jako

$$C_{gd}(S) = \frac{C_{gd}(S)}{|G| - |S|}.$$

Časová složitost zjištění skupinové degree centrality je $\mathcal{O}(m)$, kde m je počet hran v grafu, vyplývá to z algoritmu. který je popsán později v této sekci.

Degree centrality je použita téměř při každém pokusu o určení nejvlivnější osoby na sociálních sítích [22]. Zde se jedná téměř výhradně o vstupní degree centrality. Například u sociální sítě Twitter se můžeme bavit o počtu sledujících daného profilu. Zde platí, že čím více má profil sledujících, tím se považuje za vlivnější. Dále se také degree centrality využívá k odhalení podvodníků od legitimních uživatelů na online aukcích. Podvodníci mívají většinou větší degree centrality než běžní uživatelé, protože mají tendenci se domlouvat mezi sebou na umělém zvýšení ceny prodejních položek [22].

Tato míra ale například není vhodná pro hledání vrcholů vhodných pro řízení toku informací skrz sítí. Pokud uvedeme příklad tak může být vrchol

Obrázek 2.1: Degree centrality a kontrola toku sítě

v části, kde je poměrně hustý výskyt vrcholů a může mít například tvar hvězdy. Ale v grafu může být daleko větší skupina vrcholů, ke které má daný vrchol daleko, a tudíž není až tak optimálně zvolen jako vhodný pro tento účel.

Jak vidíme na Obrázku 2.1, tak vrchol 1 by měl největší degree centrality, konkétně neznormalizovaně 4, ale většina vrcholu se nachází na pravé straně grafu, takže by vrchol 1 nebyl nejvhodnějším kandidátem na vrchol pro kontrolu toku sítě.

Výpočet pro jednotlivou skupinu tedy provedeme následovně:

1. Připravíme si prázdnou množinu.

2. Podíváme se na všechny vrcholy odpovídající skupiny a všechny sousedy tohoto vrcholu, které mají jinou skupinu než tento vrchol, přidáme do připravené množiny.

3. Velikost množiny je naše finální skupinová degree centrality.

Jelikož všechny vyhovující sousedy přidáváme do množiny, tak se nám nemohou objevit duplikáty, neboť množina je abstraktní datový typ, který z definice obsahuje unikátní prvky a stejně tak je implementována množina v Pythonu [23].

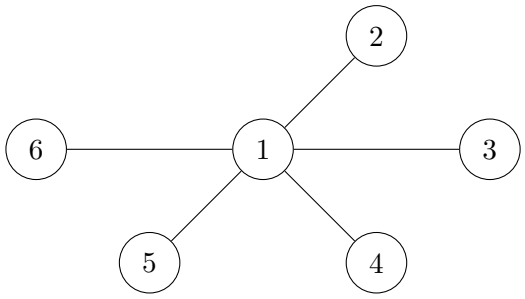
Normalizaci této míry provede tak, že získanou velikost množiny vydělíme počtem vrcholů, které nenáleží naší měřené skupině.

2.1.2 Closeness centrality

Closeness centrality detekuje v síti vrcholy, které jsou schopny velmi efektivně šířit informaci [24]. Tato míra využívá jako hlavní vlastnost vrcholu jeho vzdálenost od ostatních vrcholu grafu. Klasický closeness centrality situována na jednotlivé vrcholy bere průměrnou vzdálenost od ostatních vrcholů. Toto číslo by ovšem znamenalo, že čím menší výsledná míra je, tím je vrchol brán jako vhodnější. Aby se zachovaly vlastnosti jako u ostatních měr, tak jako výsledek bereme převrácenou hodnotu vzdálenosti od ostatních vrcholů. Tedy vzorec pro klasickou closeness centrality je

$$C_c(v) = \frac{1}{\sum_{u \in V} \mathrm{dist}(v, u)}.$$

Obrázek 2.2: Closeness centrality, hvězda



Ale běžnější je normalizovaná forma, která bere průměrnou vzdálenost od všech ostatních vrcholů, tedy

$$C_c(v) = \frac{|G| - 1}{\sum_{u \in V} \mathrm{dist}(v, u)}.$$

Jedná z podmínek je, že graf, na kterém se closeness centrality měří, musí být souvislý, pokud by nebyl, tak by existovala dvojice vrcholů, jejichž vzdálenost by byla ∞.

Na Obrázku 2.2 vidíme graf, kterému se říká *hvězda*. Ta se vyznačuje jedním centrálním vrcholem, na který jsou ostatní vrcholy napojeny. Centrálním prvek je zde vrchol 1. Ten bude mít maximální možnou closeness centrality, jelikož jeho vzdálenost od každého dalšího vrcholu je 1.

Zde je rozšíření na skupinovou closeness centrality výpočetně složitější než jak tomu bylo u skupinové degree centrality. Abychom pro skupinu U spočítali její skupinovou closeness centrality, musíme nejprve spočítat vzdálenost od všech vrcholů, které se v této skupině nenachází. Vzdálenost vrcholu od skupiny bereme jako nejmenší vzdálenost mezi vrcholem a libovolným vrcholem ze skupiny. Tudíž musíme spočítat vzdálenost pro všechny vrcholy ze skupiny od tohoto vrcholu a vybrat tu nejmenší. Poté když máme napočítané vzdálenosti od všech vrcholů, které skupině nenáleží, opět s normalizovanou formou vypočítáme normalizovanou skupinovou closeness centrality pro skupinu U jako

$$C_{gc}(U) = \frac{|G \setminus U|}{\sum_{v \in V \wedge v \notin U} \mathrm{dist}(U, v)}.$$

Skupinová closeness centrality má různé modifikace, jako vzdálenost vrcholu od skupiny se dá místo minima také použít průměr nebo maximum, ovšem nejpoužívanější variantou zůstává stále minimální vzdálenost, u které se ignoruje vnitřní struktura skupiny [5]. Většinou se totiž předpokládá, že vrcholy ve skupině mají vyšší soudržnost a komunikace mezi nimi bývá lepší než s vrcholy mimo skupinu. Toto nemusí být ovšem pravidlo, proto jsou zde i tyto 2 další varianty. Dále také existuje modifikace nazývaná closeness centrality náhodné procházky. Náhodná procházka je v matematice a fyzice užívaná formalizace intuitivní myšlenky provádění náhodných kroků. Každý další krok,

obvykle stejné délky, je učiněn náhodným směrem [25]. Tato varianta closeness centrality vznikla, jelikož informace se v síti nešíří vždy nejkratší cestou ale někdy spíše náhodně.

Closeness centrality se obecně využívá na vyhledání takových vrcholů v sítí, které mohou být ve výhodné pozici pro získávání a šíření informací napříč sítí. Tohoto se například využívá i v hledání hlavních postav zločineckých organizací. Dále se také closeness centrality využívá ke zjištění důležitosti pojmů v dokumentu. Prvně se dokument pomocí extrakce kličových frází převede na graf a poté se pomocí closeness centrality měří ona důležitost jednotlivých pojmů [24].

Algoritmus na výpočet skupinové closeness centrality tedy vypadá takto:

1. Pomocí algoritmu BFS si pro všechny vrcholy skupiny sestavíme tabulku vzdáleností od ostatních vrcholů.

2. Připravíme si nové pole, které bude mít velikost rovnou počtu prvků, které nenáleží skupině.

3. Projedeme tabulku vzdáleností a pro každý koncový vrchol vybereme nejmenší vzdálenost, která se v daném sloupci nachází a zapíšeme ji do pole.

4. Sečteme všechny prvky pole.

5. Vezmeme převrácenou hodnotu této sumy a toto je naše skupinová closeness centrality.

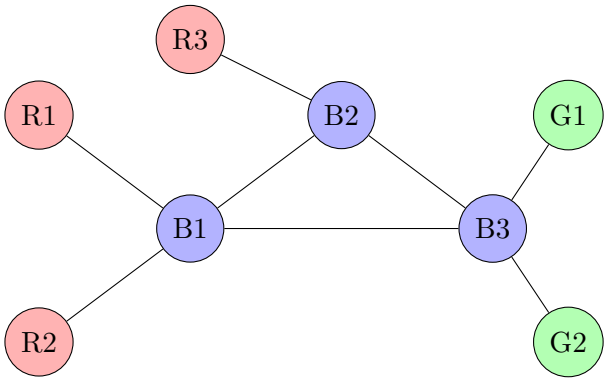
Normalizaci jak jsme si již řekli provedeme tak, že pokud máme sumu vzdáleností, tak tuto sumu vydělíme počtem vrcholů, které nenáleží skupině, jejíž centralitu počítáme, a převrácená hodnota tohoto podílu je naší normalizovanou skupinovou closeness centrality.

2.1.3 Betweenness centrality

Pro jednotlivé vrcholy je *betweenness centrality* míra, která se snaží určit důležitost vrchol v rámci polohy na cestách ostatních dvojic vrcholů. Slouží k detekci vrcholů, které mají největší kontrolu nad tokem informací v síti [26]. Individuální betweenness centrality je definována jako podíl nejkratších cest dvojic ostatních vrcholů procházejících daným vrcholem, značíme $\sigma_{xy}(v)$ ku všem nejkratším cestám dvojic ostatních vrcholů, neboli

$$C_b(v) = \sum_{x \neq y \neq v} \frac{\sigma_{xy}(v)}{\sigma_{xy}}$$

Rozšíření na skupinovou betweenness centrality může být provedeno dvěma způsoby. Zaprvé to může být bráno jako počet nejkratších cest dvojic vrcholů



Obrázek 2.3: Skupinová betweenness centrality, ukázka

mimo skupinu, které prochází alespoň jedním vrcholem, který skupině náleží a nebo to může být zavedeno jako počet nejkratších cest dvojic skupinů mimo vrcholů, které prochází všemi vrcholy dané skupiny. Tedy skupinovou betweenness centrality pro skupinu U vypočítáme jako

$$C_{gb}(U) = \sum_{u<v} \frac{\sigma_{u,v}(U)}{\sigma_{u,v}} \quad u, v \notin U.$$

První způsob je označovám jako skupinová betweenness centrality a druhý je v literatuře označován jako co-betweenness centrality [5]. My se budeme zaměřovat pouze na skupinovou betweenness centrality. Podobně jako u closeness centrality, také zde existuje varianta na náhodné procházky. Modifikace tady zní tedy na kolika náhodných cestách mezi dvojicemi ostatních vrcholů leží daný vrchol.

Na Obrázku 2.3 má například modrá skupina $B = \{B1, B2, B3\}$ maximální skupinovou betweenness centrality, neboť leží na všech cestách dvojic ostatních vrcholů.

Meření betweenness centrality se používá u problému typu návrh sítě, za účelem maximalizování cílené vlastnosti. Betweenness centrality se používá jakožto základ pro maximalizaci toku informací nebo provozu procházející daným vrcholem nebo skupinou vrcholů. Tento provoz se snažíme maximalizovat především přidáním nevelkého počtu hran tak, abychom zvýšili betweenness centrality [27]. Například na internetových obchodech může přidání hrany mezi vrcholy znamenat přidat náš cílený produkt jako doporučení u jiného produktu.

Skupinová betweenness centrality je výpočetně nejnáročnější centrality, kterou v práci rozebíráme. Dle definice se jedná o poměr nejkratších cest vrcholů mimo skupinu, které prochází vrcholem náležícím skupině, ku počtu všech nejkratších cest vrcholů mimo skupinu. A takto budeme také počítat a výpočet provedeme podle následujících kroků.

1. Připravíme si všechny kombinace dvojice vrcholů tak, aby žádný z dané dvojice nepatřil do naší skupiny a dále inicializujme proměnnou $C_{gb} = 0$
2. Průběžné výsledky budeme ukládat do matice, jejíž souřadnice budou tvořit prvky dvojice, označme M_{uv} pro vybranou dvojici vrcholů u, v a druhou matici stejných rozměrů MR_{uv}
3. Vybereme dosud nepoužitou dvojici a spočteme počet nejkratších cest a délku této cesty mezi vrcholy vybrané dvojice modifikací Algortimu BFS 1, označíme p_n, d_n a uložíme do M_{uv}
4. Bod **2.** a **3.** opakujeme pro všechny kombinace dvojic vrcholů
5. Odstraníme všechny hrany takové, které mají alespoň 1 koncový vrchol náležící skupině
6. Provedeme body **2.** a **3.** znovu s tímto grafem a výsledky, ale zapisujeme do MR_{uv}, tyto výsledky označujme jako p_r a d_r
7. Nyní porovnáme prvky obou matic dle jednotlivých souřadnic následovně
   - **7.1.** Pokud $d_n \neq d_r$ pak bez dané skupiny neexistuje cesta stejné délky [2] nebo neexistuje vůbec, tedy $C_gb+ = 1$
   - **7.2.** Pokud $d_r = d_n$, pak jsme našli nějaké cesty, tedy $C_{gb}+ = \frac{p_n - p_r}{p_n}$
8. C_{gb} je skupinová míra centrality dané skupiny

U této míry se může stát, že míra pro danou skupinu nelze normalizovat, toto se děje v případě, že skupinu tvoří pouze 1 prvek. V ostatních případech je normalizace pro graf G a skupinu U dána tímto vzorcem

$$NC_{gb} = \frac{C_{gb}}{(|G| - |U|) \cdot (|G| - |U| - 1)}.$$

[2] Pokud odebereme hrany z grafu, nemůže se nám délka nejkratší cesty zmenšit

Kapitola 3

Měření

3.1 Představení datasetů

Všechny datasety pocházejí z laboratoře Stanfordské univerzity SNAP [3]. Jedná se o knihovnu datasetů, které jsou používány pro výzkumné aktivity na sociálních a informačních sítích.

3.1.1 Facebook dataset

Tento dataset je z jedné z nejznámějších sociálních sítí a to Facebook. Zde jsou vrcholy brány jako ověřené profily různých ověřených profilů známých osobností nebo společností a hrany jsou zde jako vzájemné označení "To se mi líbí"mezi jednotlivými profily. Tento dataset byl sestaven v listopadu roku 2017 přes Facebook Graph API. Jednotlivé profily jsou zde rozděleny na politické osobnosti, vládní organizace, televizní společnosti a firmy. [28] Podle těchto rozdělení budeme také uvažovat jednotlivé skupiny v rámci měření našich skupinových centralit. Tento graf je neorientovaný. Dále také tento graf má 22 470 vrcholů a 171 002 hran.

3.1.2 GitHub dataset

GitHub je webová služba podporující vývoj softwaru za pomoci verzovacího nástroje Git. Tato služba momentálně hostuje více než 200 miilionu repozitářu a pro uživatele poskytuje funkce sociálních sítí [4]. Dataset vzniknul v červnu 2019, přes veřejné API této služby. Vrcholy v tomto grafu jsou vývojáři, kteří mají alespoň 10 repozitářů a hrany mezi vrcholy jsou vzájemné sledování profilů na GitHubu. Rozdělení je zde pouze na 2 skupiny a to jestli je uživatel webový vývojář nebo znalostní inženýr. Toto bylo získáno z informací na profilu jednotlivých uživatelů. Graf obsahuje 37 700 vrcholů a 289 003 hran.

[3] https://snap.stanford.edu/index.html
[4] https://github.com/about

3.1.3 Eu-core dataset

Tento dataset byl získán z emailových dat větších výzkumných organizací v Evropě [29]. V tomto grafu máme 1005 vrcholů reprezentující jednotlivé výzkumné pracovníky a hrany zde jsou za podmínky, že si uživatelé mezi sebou poslali alespoň 1 email, těch je 25 571. Jednotliví pracovnící patří do jednoho ze 42 oddělení a podle toho jsou také rozděleny na jednotlivé skupiny.

3.1.4 LastFM dataset

LastFM je internetové rádio a systém pro doporučování hudby [30]. Tato služba také umožňuje svým uživatelům si vytvořit profil a dále také sledovat profily ostatních uživatelů, čemuž má tato služba také prvky sociální sítě. Náš dataset byl vytvořen v březnu roku 2020 z veřejného API. Tento dataset má 7 624 vrcholů, které jsou zde jako jednotlivý uživatelé asijských zemí a 27 806 hran, které zde reprezentují již zmiňované sledování mezi uživateli [31] . Vrcholy jsou zde rozděleny dle národnosti jednotlivých uživatelů.

3.1.5 Twitch dataset

Twitch je nejpopulárnější platforma pro živé vysílání, která byla původně zaměřeno na videohry, nicméně se postupem času rozrostla a vysílání hudby, talkshow nebo umění [32]. Náš dataset byl sestaven na jaře roku 2018 opět z veřejného API této služby. Vrcholy jsou zde jako jednotliví tvůrci na této službě a hrany tohoto grafu jsou zde jako vzájemné sledování mezi profily uživatelů [33]. Jednotlivé profily mají v datasetu vícero atributů, ovšem na naši úlohu se nejvíce hodí atrbitut rozdělující uživatele podle jazyku, ve kterém tvoří svůj obsah, těch je celkem v datasetu 21. Graf obsahuje 168 114 vrcholů a 6 797 557 hran.

3.1.6 Souhrn

Jak již bylo zmíněno, datasety bylo nutno vzorkovat, ani s nejlepšími výpočetní algoritmy ke zjištění skupinových měr centralit by nebylo s naším vybavením možné dojít k výsledkům v rozumné době. K vzorkování bylo použito vzorkování náhodnou procházkou vysvětleno v sekci 1.4.1. U datasetů Eu-Core a LastFM k vzorkování dojít nemuselo, jelikož datasety obsahují rozumný počet vrcholů.

	Vrcholů	Hran	Skupin	Vrcholů po vzorkování
Facebook	22 470	171 002	4	2 616
GitHub	37 700	289 003	2	2 176
Eu-Core	1 005	25 571	42	1 005
LastFM	7 624	27 806	19	7 624
Twitch	168 114	6 797 557	21	4 347

Tabulka 3.1: Souhrn metrik datasetů

3.2 Představení testovacího prostředí

3.2.1 Software

Implementace skupinových měr centralit byla provedena v jazyce Python. Python je vysokoúrovňový interpretovaný jazyk velice vhodný na práci s daty a různými statistickými metodami [5]. Konkrétně byla implementace provedena v Jupyter notebooku, což je označení pro skupinu softwarových produktů, které zpřístupňují programování pomocí webového rozhraní [34] . K vizualizaci grafů nám pomůžou knihovny pro jazyk Python Networkx [6] a také Matplotlib [7].

3.2.2 Spouštení Jupyter notebooku

1. Nejprve si musíme stáhnout distribuci Pythonu, nejjednodušší způsob je přes nástroj Anaconda [8]. Tu potom musíme ještě nainstalovat.

2. Ke spuštení Jupyter notebooku stačí do Anaconda terminálu napsat `jupyter notebook`. Ve webovém prohlížeči by se nám měla otevřít nová záložka s domácím adresářem našeho filesystému.

3. Ve webovém rozhraní stačí pouze najít naší složku a spustit soubor `prakticka_cast_bp_turekmat.ipynb`.

3.2.3 Použité třídy

Pro udržení si informací o daném grafu budeme využívat 2 třídy. První z nich bude třída `Node` neboli vrchol. Tato třída si bude držet informace o konkrétní instanci vrcholu, jeho zařezení do skupiny a o daných sousedech tohoto vrcholu. Budeme také si u každého vrcholu držet uměle vytvořenou proměnnou `ID`, pro lepší identifikaci vrcholů. Druhou třídou bude třída `Graph`. Tato třída bude sdružovat a ukládat jednotlivé instance třídy `Node` tedy vrcholy.

[5] https://www.python.org/about/
[6] https://networkx.org/
[7] https://matplotlib.org/
[8] https://www.anaconda.com/download/

Bude si také držet informace o počtech, takže například počet vrcholů a počet vrcholů nacházejících se v jedné skupině, to nám poté usnadní výpočty skupinových měr centralit.

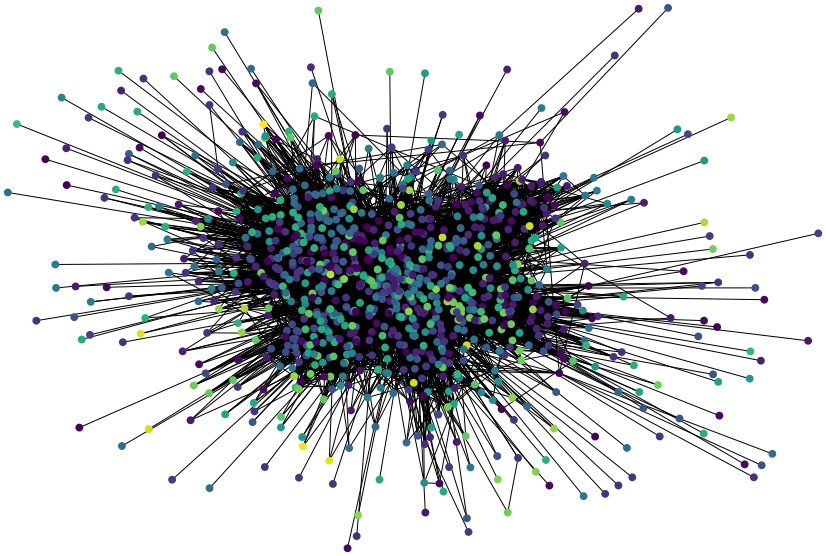
3.3 Měření míry centrality na datasetech

V této sekci si postupně rozebereme měření skupinových měr centralit na jednotlivých datasetech. V rámci jednotlivého měření bude vždy také vykreslen graf znázorňující daný dataset. Podle Python knihovny NetworkX jsme vykreslili graf znázorňující dataset a podle také této knihovny jsme ověřili správnost výsledků naší implementace. Z tohoto grafu se budeme snažit predikovat různé míry a dále podle skutečně naměřených hodnot tyto predikce vyhodnotíme a prodiskutujeme dané výsledky.

3.3.1 Měření na datasetu Eu-core

Jak jsme si již řekli v představení datasetu, tak dataset Eu-core má 42 skupin a tudíž snažit se v Obrázku 3.1 hledat komplexnější predikce bude složitější. Čeho si můžeme všimnout, tak že nejpočetnější skupinou bude nejpíše skupina tmavě modrých vrcholů, které znázorňují skupinu 4. Tyto vrcholy vypadají poměrně soustředěné ve středu grafu, to by mohlo naznačovat vysoké hodnoty u skupinové closeness centrality.

Obrázek 3.1: Graf datasetu Eu-core



Tabulka 3.2: Měření Eu-core

ID Skupiny	Skupinová betweenness centrality	Skupinová closeness centrality	Skupinová degree centrality	Podíl vrcholů v grafu	Barva v grafu
0	0.05293	0.09113	0.52772	0.04776	Čokoládová
1	0.04123	0.07629	0.58026	0.06368	Červená
2	0.00682	0.01021	0.24078	0.00995	Bledě modrá
3	0.0027	0.00291	0.15385	0.01095	Okrová
4	0.14388	0.16121	0.71982	0.10746	Tmavě modrá
5	0.00838	0.00806	0.28306	0.01791	Růžová
6	0.02384	0.03022	0.40833	0.02587	Hluboká růžová
7	0.04151	0.07464	0.49305	0.05075	Tyrkysová
8	0.04006	0.05527	0.41468	0.01891	Světle korálová
9	0.02082	0.03951	0.41466	0.03085	Růžová
10	0.03568	0.04816	0.53376	0.03782	Olivová
11	0.00834	0.02068	0.26646	0.02886	Azurová
12	0.00251	0.00159	0.08138	0.00299	Měděná
13	0.03198	0.04364	0.46979	0.02587	Lesní zelená
14	0.13154	0.18136	0.69643	0.08955	Žlutá
15	0.03634	0.06442	0.5284	0.05224	Královská modrá
16	0.00716	0.03613	0.39126	0.02488	Mořská zelená
17	0.04554	0.05718	0.5	0.03383	Šedá
18	0	0.0	0.00305	0.00099	Tmavě olivová
19	0.01512	0.03394	0.37931	0.02886	Lososová
20	0.03122	0.03237	0.36831	0.01393	Indigo
21	0.05809	0.09352	0.59223	0.05871	Modrá
22	0.00549	0.00838	0.28586	0.02388	Khaki
23	0.0135	0.01371	0.34202	0.02687	Tmavě růžová
24	0.01944	0.01376	0.18776	0.00597	Dodger modrá
25	0.02811	0.0458	0.28163	0.00597	Zelená
26	0.01229	0.01724	0.25589	0.00896	Tmavě oranžová
27	0.00025	0.0107	0.23258	0.00995	Oranžová
Pokračování na další straně					

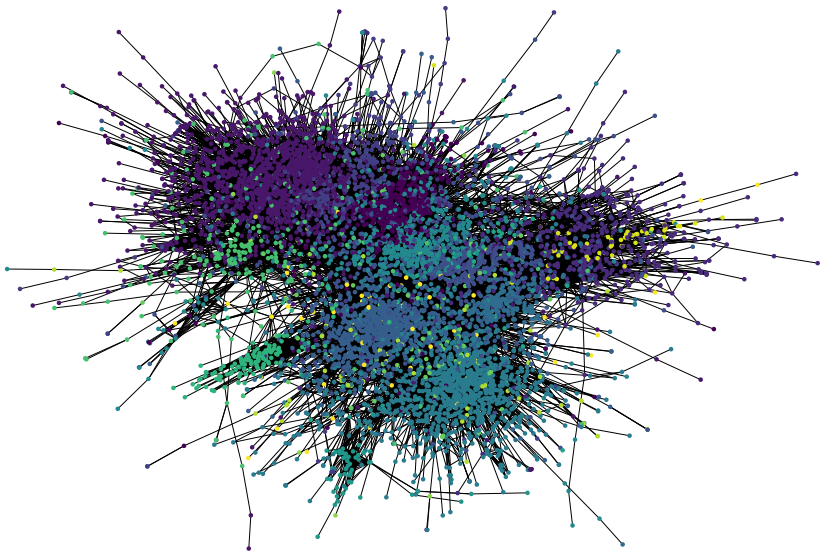
Table 3.2 Pokračování z předešlé strany

ID Skupiny	Skupinová betweenness centrality	Skupinová closeness centrality	Skupinová degree centrality	Podíl vrcholů v grafu	Barva v grafu
28	0.00088	0.00313	0.11554	0.00796	Tmavě modrozelená
29	0.00107	0.00073	0.06212	0.00398	Tmavě fialová
30	0.00023	0.00692	0.08554	0.00398	Bronzová
31	0.00767	0.01193	0.21166	0.00796	Rajčatová
32	0.00486	0.00646	0.16479	0.00896	Indiánská červená
33	0	1e-05	0.00609	0.00099	Bílá
34	0.05778	0.07463	0.46865	0.01294	Hnědá
35	0.00866	0.00708	0.21891	0.01294	Ocelová modrá
36	0.08634	0.08568	0.58817	0.02189	Tmavě červená
37	0.01545	0.01743	0.31276	0.01393	Navy
38	0.01307	0.01476	0.26899	0.01194	Karmínová
39	0	0.00011	0.02848	0.00299	Orchid
40	0.0013	0.00192	0.12424	0.00398	Tmavě tyrkysová
41	0	1e-05	0.00508	0.00199	Limetková zelená

Podle naměřených hodnot v Tabulce 3.2 vidíme, že opravdu skupina 4 má nejvíce vrcholů v grafu. Co se týče naší predikce ohledně skupinové closeness centrality této skupiny, tak v rámci grafu má druhou nejvyšší. Když se podíváme na naměřené hodnoty, tak tato skupina má také nejvyšší skupinovou degree centrality a i skupinovou betweeness centrality.

3.3.2 Měření na datasetu LastFM

Obrázek 3.2: Graf datasetu LastFM



Zde na Obrázku 3.2 máme vykreslený graf datasetu ze sítě LastFM. Dataset LastFM pro zopakování obsahuje 19 různých skupin ovšem na 3.2 vidíme nejčastěji fialové, modré a tyrkysové, které náleží se skupinám s číslem 8, 17 a 0 v tomto pořadí.

Můžeme si všimnout, že napříč středem grafu prochází žlutá skupina, tedy skupina 5 a tak by mohla mít nejspíše vysokou skupinovou closeness centrality. Dále také vidíme, že modré a fialové vrcholy jsou rozprostřeny napříč celým grafem, to by mohlo také znamenat, že jsou sousedy více vrcholů jiných skupin, což by jim také zvedlo skupinovou degree centrality.

Nyní si rozebereme naměřenené hodnoty dle Tabulky 3.3. Naše predikce byla z části úspěšná, když se podíváme se míry pro skupinovou degree centrality, můžeme si všimnout, že opravdu nejlepších výsledků dosahuje skupina 6, tedy skupina s tmavě modrými vrcholy a hned druhou v pořadí je skupina 17. Co se týče skupinové closeness centrality, tak nejlepších výsledků dosahuje opět skupina 6 a 17, dále pak skupina 14. Skupinovou betweeness centrality má nejvyšší opět skupina 17 a skupina 0 s červenými vrcholy. Skupina 6, která u ostatních skupinových měr centralit měla nejvyšší výsledky má skupinovou betweenness centrality 0.08279 a v pořadí je 5. nejvyšší.

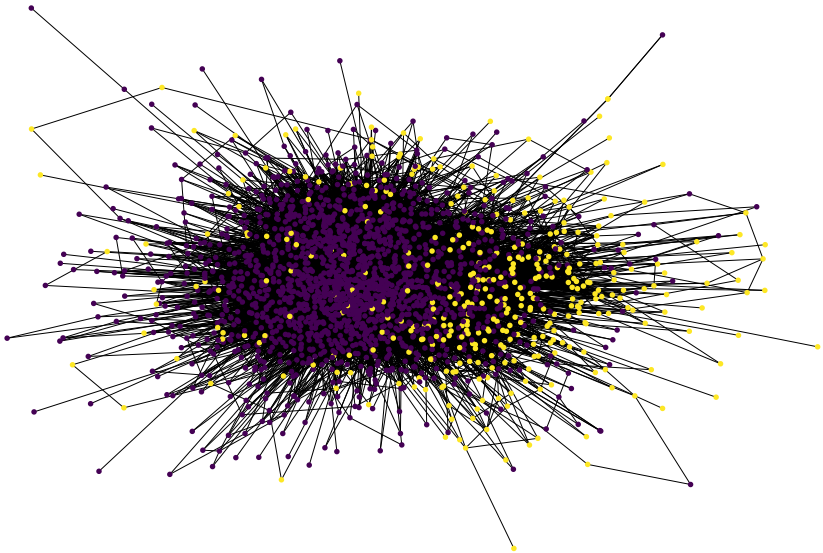
Tedy u datasetu LastFM vidíme, že skupinou, která měla napříč všemi třemi skupinovými centralitami nejvyšší výsledky byla skupina 17. Tato skupina má také největší počet vrcholů.

ID Skupiny	Skupinová betweenness centrality	Skupinová closeness centrality	Skupinová degree centrality	Podíl vrcholů v grafu	Barva v grafu
0	0.16084	0.37407	0.08229	0.14402	Tyrkysová
1	0.00465	0.2529	0.00489	0.00708	Korálová
2	0.02358	0.28156	0.00993	0.00958	Magenta
3	0.0327	0.33409	0.02799	0.06755	Zelená
4	0.00096	0.249	0.00289	0.0021	Modrozelená
5	0.05587	0.34346	0.03747	0.05129	Žlutá
6	0.08279	0.4048	0.1069	0.08591	Tmavě modrá
7	0.0152	0.30404	0.0187	0.01076	Navy
8	0.11973	0.37176	0.05492	0.06139	Fialová
9	0.01356	0.28633	0.00925	0.00761	Khaki
10	0.07841	0.36769	0.06581	0.17091	Šedá
11	0.00155	0.27533	0.00494	0.0181	Olivová
12	0.00259	0.27533	0.00515	0.00748	Královská modrá
13	0.00085	0.26011	0.00198	0.00826	Světle zelená
14	0.10432	0.37465	0.08038	0.07476	Hnědá
15	0.03068	0.31533	0.0224	0.03371	Růžová
16	0.03446	0.34329	0.0403	0.03332	Tmavě červená
17	0.17253	0.39683	0.09865	0.20619	Modrá

Tabulka 3.3: Měření LastFM

3.3.3 Měření GIT

Obrázek 3.3: Graf datasetu GIT



Jak jsme si již řekli, u datasetu dělíme vrcholy pouze na 2 skupiny a to na skupinu webových inženýrů a skupinu znalostním inženýrů. Již z Obrázku 3.3 vidíme, že grafu naprosto dominuje fialová barva, ta reprezentuje webové inženýry. Je nejspíš snadné odhadnout, že fialové vrcholy budou sousední větší části vrcholů žlutým, než tomu bude naopak. I u skupinové betweenness centrality by šlo předpovídat, že na cestě mezi dvěma vrcholy žluté skupiny bude pravděpodobněji i fialový vrchol, než v opačném případě.

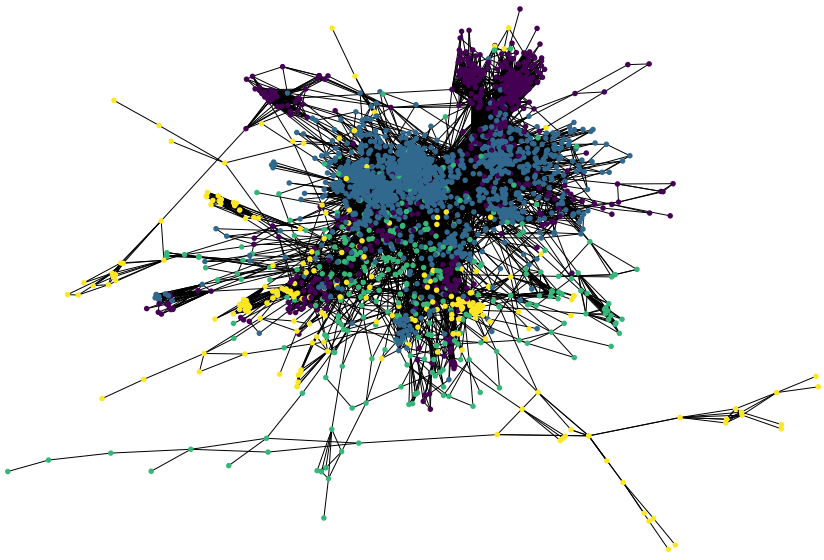
Naměřené hodnoty potvrzují naší domněnku. Skupina 0, tedy webových inženýru, na Obrázku 3.3 znázorněna vrcholy fialové barvy má u všech tří skupinových centralit vysoké normalizované skore. U skupinové betweenness centrality má dokonce výsledek velice blízký k číslu 1. Skupina 1 má velmi nízkou skupinovou betweenness centrality, ale výsledky u skupinové closeness a skupinové degree centrality už má poměrně vyšší vzhledem k počtu vrcholů náležící této skupině.

ID Skupiny	Skupinová betweenness centrality	Skupinová closeness centrality	Skupinová degree centrality	Podíl vrcholů v grafu	Barva v grafu
0	0.97458	0.93717	0.94134	0.83548	Fialová
1	0.03418	0.63902	0.45105	0.16452	Žlutá

Tabulka 3.4: Měření GIT

3.3.4 Měření Facebook

Obrázek 3.4: Graf datasetu Facebook



Náš dataset ze sítě Facebook, jak jsme si již řekli, se skládá ze 4 skupin. Podle toho vidíme na Obrázku 3.4 také rozdělení 4 různé barvy. Vidíme, že pravdě-

ID Skupiny	Skupinová betweenness centrality	Skupinová closeness centrality	Skupinová degree centrality	Podíl vrcholů v grafu	Barva v grafu
Firmy	0.20182	0.52574	0.25677	0.15291	Fialová
Televizní společnosti	0.04577	0.45605	0.12308	0.12729	Zelená
Vládní organizace	0.44584	0.54892	0.3824	0.48318	Modrá
Politické osobnosti	0.14699	0.50088	0.25638	0.23662	Žlutá

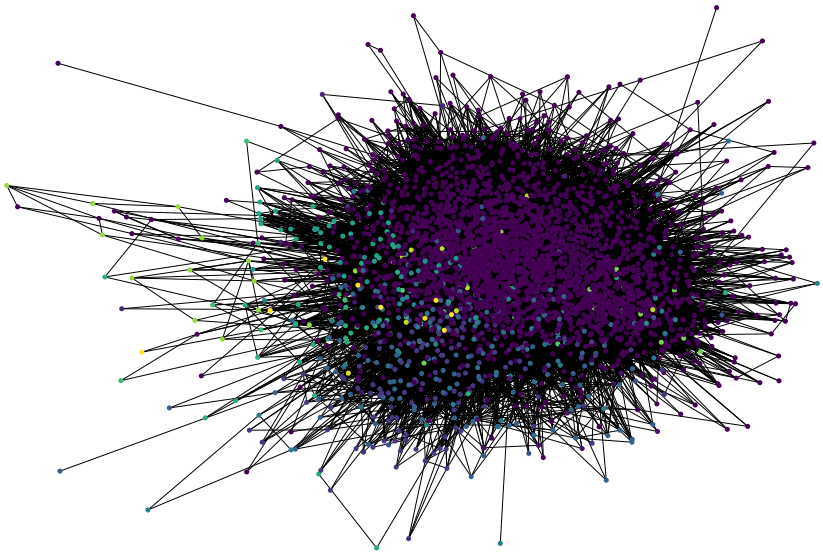
Tabulka 3.5: Měření Facebook

podobně nejvíce vrcholů bude mít skupina vrcholů modré barvy, to je skupina vládních organizací. Jak už jsme zjistili z předešlých měření, skupina, která s největším počtem vrcholů obvykle dosahuje vyšších výsledků u skupinových měr centralit. Dále to vypadá, že by žlutá skupina, skupina politických osobností, mohla mít vysokou skupinovou degree centrality. U 3.4 je také zajímavé rozložení politických osobností, jak je vidno z grafu, tak se tvoří poměrně viditělné shluky v rámci této skupiny. To může být zapřičiněno politickou příslušností jednotlivých osobností.

Jak vidíme z Tabulky 3.5, tak domněnky o skupině vládních organizací byly oprávněné, ve všech třech skupinových mírách dosáhly nejvyšších výsledků v rámci tohoto grafu. U tohoto měření je také zajímavé, že skupina firem a politických osobností má téměř shodnou skupinovou degree centrality, přestože skupina firem má daleko méně zastoupených vrcholů v tomto grafu. Toto je pravděpodobně dáno tím, že politické osobnosti mají hrany v grafu, tedy "To se mi líbí"na sociální síti, z větší části pouze mezi sebou, zatímco skupina firem má tento vztah spíše s ostatními skupinami.

3.3.5 Měření Twitch

Obrázek 3.5: Graf datasetu Twitch



Graf Twitche je složený ze skupin, dle jazyku, ve kterém tvůrci vytváří obsah. Jako první predikce by mohla být, že drtivá většina obsahu přes tuto službu bude vytvářena v angličtině. Pro tvůrce je vhodná tím, že bude mít největší dosah a tím pádem by měla mít i daleko vyšší naměřené hodnoty skupinových měr centralit. Jak vidíme dle Obrázku 3.5 je tomu tak, vrcholy fialové barvy zde dominují. Zároveň je velmi pravděpodobné, že 2 tvůrce s naprosto jiným jazykem tvorby, bude spojovat uživatel, který bude mluvit anglicky, tím pádem by také skupinová betweenness centrality měla být vysoká. Naopak u ostatních skupin bychom neměli předpokládat vysoké hodnoty skupinové betweenness centrality.

Z Tabulky 3.6 vidíme, že opravdu skupina anglických tvůrců má s velikým náskokem nejvyšší naměřené hodnoty. Za povšimnutí dále taky stojí skupiny a tvůrci "DE", tedy němečtí, "ES", španělští, a "FR", francouzští, tyto skupiny, k poměru s počtem vrcholů v grafu, dosahují také vysokých hodnot. Naopak "RU", skupina ruských tvůrců, má hodnoty nižší než jsme očekávali, jelikož patří k jedním z nejvíce používaných jazyků světa [35].

3.3.6 Souhrn měření

Jednou z otázek po tomto měření bylo, jak spolu souvisí různé skupinové míry central mezi sebou a jak jsou závislé také na velikosti skupiny. Na tuto otázku

ID Skupiny	Skupinová betweenness centrality	Skupinová closeness centrality	Skupinová degree centrality	Podíl vrcholů v grafu	Barva v grafu
EN	0.95216	0.92	0.91304	0.85007	Fialová
SV	0.00149	0.50596	0.0892	0.00681	Modrá
OTHER	0.00051	0.49807	0.06841	0.00385	Zelená
IT	0.00023	0.48115	0.03605	0.00563	Žlutá
DE	0.00304	0.53567	0.16891	0.02993	Tmavě modrá
ES	0.0017	0.52154	0.12541	0.01719	Oranžová
KO	3e-05	0.45752	0.01961	0.00267	Růžová
FR	0.00215	0.53577	0.168	0.03704	Šedá
PT	0.00066	0.50398	0.08102	0.00533	Šedá
RU	0.00032	0.49593	0.06543	0.0083	Hnědá
NL	0.00038	0.48946	0.05077	0.00207	Tyrkysová
JA	0.00167	0.51002	0.11283	0.00474	Magenta
NO	0.00067	0.46793	0.0454	0.00148	Olivová
HU	7e-05	0.46498	0.01898	0.00089	Tmavě červená
ZH	0.00088	0.49444	0.06623	0.01126	Navy
FI	1e-05	0.44493	0.00593	0.00059	Béžová
PL	0.00115	0.4916	0.06536	0.00267	Bílá
TR	0.00052	0.47903	0.03087	0.00178	Tmavě zelená
CS	0.0003	0.47835	0.03215	0.00474	Čokoládová
DA	0.00066	0.4906	0.04814	0.00296	Indigo

Tabulka 3.6: Měření Twitch

	Skupinová betweenness centrality	Skupinová closeness centrality	Skupinová degree centrality	Podíl vrcholů v grafu
Skupinová betweenness centrality	1	0.83716	0.52914	0.42647
Skupinová closeness centrality	0.83716	1	0.85944	0.72284
Skupinová degree centrality	0.52914	0.85944	1	0.93598
Podíl vrcholů v grafu	0.42647	0.72284	0.93598	1

Tabulka 3.7: Korelace veličin

nám může odpovědět výběrový korelační koeficient. Obecně nám korelační koeficient řekne míru vzájemné lineární závislosti mezi dvěma veličinami [36]. Hodnoty byly naměřeny podle knihovny pro Python NumPy [9].

Jak vidíme z Tabulky 3.7 tak všechny veličiny jsou mezi sebou pozitivně korelované, to znamená, že pokud roste jedna, tak roste i druhá. Nejvyšší korelaci vidíme mezi skupinovou degree centrality a podílem vrcholů v grafu. To je také logické, jedině, jak by mohl růst podíl vrcholů v grafů a zároveň by skupinová degree centrality zůstala stejná je v případě, kdy další přidané vrcholy ze skupiny byly spojeny pouze s vrcholy v rámci jeho vlastní skupiny, což se v reálných případech moc neděje. Dále také stojí za zmínku poměrně vysoká korelace mezi skupinovou closeness centrality a skupinovou degree centrality. Vysvětlení pro toto by mohlo být, že s čím více vrcholy je daná skupina spojená, tím blížší bude průměrná vzdálenost k vrcholů, cizích skupin.

Z napočítaných korelací je nejnižší mezi dvojicí podílu vrcholů v grafu a skupinové betweenness centrality. Skupinová betweenness centrality spíš než na množství vrcholů závisí na výhodné pozici skupiny. Například pokud je skupina na všech cestách mezi dvěma skupinami, tak nemusí mít sama skupina hodně vrcholů aby toto bylo realizovatelné a stejně bude mít poměrně vysokou skupinovou betweenness centrality.

[9] https://numpy.org/doc/stable/index.html

Závěr

Cílem této práce bylo seznámení se s mírami centralit a jejich rozšířením na skupinové míry centralit. Dále jsme také chtěli vhodné míry implementovat a naměřit hodnoty daných skupinových měr centralit na datasetech skutečných reálných sociálních sítí a tyto výsledky prodiskutovat. Všechny tyto body se nám podařilo splnit a řádně popsat jak došlo k jejich naplnění.

V kapitole 1 jsme si zavedli vhodnou notaci a potřebné definice k zavedení a úplnému pochopení skupinových měr centralit. Taky jsme si představili algoritmy, které jsme později využili v naší implementaci.

Kapitola 2 už se věnovala mírám centralit, první jsme si obecně představili klasické míry centrality a poté naše zkoumané skupinové míry centralit. Celkově jsme se detailně podívali na 3 míry centralit a to degree centrality, closeness centrality a betweenness centrality. U všech těchto měr centralit jsme si řekli způsob, kterým se vypočítá jejich klasická varianta a uvedli jsme také algoritmy pro výpočet jejich varianty pro využití na skupiny. Dále jsme si také zmínili možnost využití a dále také další vlastnosti.

Kapitola 3 už se věnovala jednotlivým měřením. Prvně jsme si představili datasety, na kterých jsme měřili skupinové míry centrality a dále také co jsme použili za software k implementaci našich algoritmů. Poté jsme v této kapitole pro každý dataset si vykreslili graf, který tento dataset znázorňuje a snažili se udělat predikce na tomto vyobrazení. Tyto predikce jsme poté porovnali se skutečně naměřenými hodnotami a podívali se na zajímavé výsledky, které jsme také prodiskutovali.

Po naměření hodnot také vyvstala otázka, jak spolu dané skupinové míry centrality souvisí. Na tuto otázku nám odpověděl výpočet korelace mezi jednotlivými skupinovými měrami centralit. Tento výpočet jsme také prodiskutovali a snažili se vysvětlit proč a jak by tyto výsledky mohli souviset.

V budoucnu by se tato práce dala vylepšit z hlediska použitých algoritmů. Existují algoritmy, které jsou schopny dané míry vypočítat s větší efektivností nebo také existují jejich varianty, které mohou tyto míry počítat paralelně, což opět přispívá na efektivitě výpočtu [37, 38]. Dále také je možné implementovat

funkčnost hledání skupiny vrcholů, který by danou míru maximalizovala. Zde se ale domníváme, že žádný algoritmus pracující v polynomiálním čase nemůže existovat, pokud předpokládáme platnost hypotézy $P \neq NP$. Můžeme ale například hledat aproximační algoritmy.

Literatura

[1] Havlová, J.: Sociální síť. KTD: Česká terminologická databáze knihovnictví a informační vědy (TDKIV), 2003. Dostupné z: https://aleph.nkp.cz/F/?func=direct&doc_number=000015947&local_base=KTD

[2] Černá, M.; Černý, M.: Úvod do sociálních sítí: měření vlivu. Dostupné z: https://clanky.rvp.cz/clanek/k/g/15131/UVOD-DO-SOCIALNICH-SITI-MERENI-VLIVU.html

[3] Wikipedia: Centralita. Dostupné z: https://cs.wikipedia.org/wiki/Centralita

[4] Luo, J.-D.: *Social Network Analysis*. Beijing: Social Science Academic Press, druhé vydání, 2004.

[5] Everett, M.; Borgatti, S.: The Centrality of Groups and Classes. *Journal of Mathematical Sociology*, ročník 23, leden 1999: s. 181–201, doi:10.1080/0022250X.1999.9990219.

[6] Diestel, R.: *Graph Theory*. Springer Publishing Company, Incorporated, páté vydání, 2017, ISBN 3662536218.

[7] Matoušek, J.; Nešetřil, J.: *Kapitoly z diskrétní matematiky*. Karolinum, 2000, ISBN 978-80-246-1740-4.

[8] Koliha, J.: Axiomatické zavedení reálných čísel. *Pokroky matematiky, fyziky a astronomie*, 1969, ISSN 0032-2423. Dostupné z: https://dml.cz/handle/10338.dmlcz/137249

[9] Weisstein; W., E.: Indegree. Dostupné z: https://mathworld.wolfram.com/Indegree.html

[10] Knop, D.; Malík, J.; Suchý, O.; aj.: Algoritmy a grafy. Dostupné z: https://courses.fit.cvut.cz/BI-AG1/lectures/media/bi-ag1-p1-handout.pdf

[11] Richta, K.: Složitost algoritmů. Dostupné z: https://cw.fel.cvut.cz/b182/_media/courses/b6b36dsa/dsa-3-slozitostalgoritmu.pdf

[12] Arora, S.; Barak, B.: *Computational Complexity: A Modern Approach*. USA: Cambridge University Press, první vydání, 2009, ISBN 0521424267.

[13] Lawler, G. F.; Limic, V.: *Random Walk: A Modern Introduction*. Cambridge Studies in Advanced Mathematics, Cambridge University Press, 2010, doi:10.1017/CBO9780511750854.

[14] Lu, J.; Li, D.: Sampling Online Social Networks by Random Walk. In *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*, HotSocial '12, New York, NY, USA: Association for Computing Machinery, 2012, ISBN 9781450315494, str. 33–40, doi:10.1145/2392622.2392628. Dostupné z: https://doi.org/10.1145/2392622.2392628

[15] Hardiman, S. J.; Katzir, L.: Estimating Clustering Coefficients and Size of Social Networks via Random Walk. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, New York, NY, USA: Association for Computing Machinery, 2013, ISBN 9781450320351, str. 539–550, doi:10.1145/2488388.2488436. Dostupné z: https://doi.org/10.1145/2488388.2488436

[16] 2022, C. I.: Social network analysis 101: centrality measures explained. Dostupné z: https://cambridge-intelligence.com/keylines-faqs-social-network-analysis/

[17] Bourjolly, J.-M.; Laporte, G.; Pesant, G.: Heuristics for finding k-clubs in an undirected graph. *Computers & Operations Research*, ročník 27, č. 6, 2000: s. 559–569, ISSN 0305-0548, doi:https://doi.org/10.1016/S0305-0548(99)00047-7. Dostupné z: https://www.sciencedirect.com/science/article/pii/S0305054899000477

[18] Friggeri, A.; Fleury, E.: Maximizing the Cohesion is NP-hard. 09 2011.

[19] Marx, D.: Graph colouring problems and their applications in scheduling. *Periodica Polytechnica Electrical Engineering*, ročník 48, č. 1-2, 2004. Dostupné z: https://pp.bme.hu/ee/article/view/926

[20] Karger, D.; Motwani, R.; Ramkumar, G. D. S.: Improved approximation algorithms for MAXk-CUT and MAX BISECTION. 1993. Dostupné z: https://doi.org/10.1007/BF02523688

[21] Langville, A. N.; Meyer, C. D.: *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, 2006, ISBN 9780691152660. Dostupné z: http://www.jstor.org/stable/j.ctt7t8z9

[22] Neo4J: Degree Centrality. Dostupné z: https://neo4j.com/docs/graph-data-science/current/algorithms/degree-centrality/

[23] Foundation, P. S.: Sets. Dostupné z: https://docs.python.org/2/library/sets.html

[24] NEO4J: Closeness Centrality. Dostupné z: https://neo4j.com/docs/graph-data-science/current/algorithms/closeness-centrality/

[25] Aldous, D.; Fill, J. A.: Reversible Markov Chains and Random Walks on Graphs. 2002, unfinished monograph, recompiled 2014, available at http://www.stat.berkeley.edu/$\sim$aldous/RWG/book.html.

[26] Neo4J: Betweeness Centrality. Dostupné z: https://neo4j.com/docs/graph-data-science/current/algorithms/betweenness-centrality/

[27] Freeman, L. C.: Centrality in social networks conceptual clarification. *Social Networks*, ročník 1, č. 3, 1978: s. 215–239, ISSN 0378-8733, doi:https://doi.org/10.1016/0378-8733(78)90021-7. Dostupné z: https://www.sciencedirect.com/science/article/pii/0378873378900217

[28] Rozemberczki, B.; Allen, C.; Sarkar, R.: Multi-scale Attributed Node Embedding. Dostupné z: https://snap.stanford.edu/data/facebook-large-page-page-network.html

[29] Leskovec, J.; Kleinberg, J.; Faloutsos, C.: Characteristic Functions on Graphs: Birds of a Feather, from Statistical Descriptors to Parametric Models. Dostupné z: https://snap.stanford.edu/data/email-Eu-core.html

[30] Wikipedia: Last.fm. Dostupné z: https://cs.wikipedia.org/wiki/Last.fm

[31] Rozemberczki, B.; Sarkar, R.: Graph Evolution: Densification and Shrinking Diameters. Dostupné z: https://snap.stanford.edu/data/feather-lastfm-social.html

[32] Elise, A.: Twitch Hits Over 45 Million Monthly Viewers. Dostupné z: https://www.ibtimes.com/twitch-hits-over-45-million-monthly-viewers-1543751)

[33] Rozemberczki, B.; Sarkar, R.: Twitch Gamers: a Dataset for Evaluating Proximity Preserving and Structural Role-based Node Embeddings. Dostupné z: https://snap.stanford.edu/data/twitch_gamers.html

[34] Tišnovský, P.: Interpret programovacího jazyka Clojure integrovaný do Jupyter Notebooku. Dostupné z: https://www.root.cz/clanky/interpret-programovaciho-jazyka-clojure-integrovany-do-jupyter-notebooku/

[35] Julian, G.: What are the Most Spoken Languages in the World? Dostupné z: http://tony-silva.com/eslefl/miscstudent/downloadpagearticles/mostspokenlangs-fluentin3months.pdf

[36] Asuero, A. G.; Sayago, A.; González, A. G.: The Correlation Coefficient: An Overview. *Critical Reviews in Analytical Chemistry*, ročník 36, č. 1, 2006: s. 41–59, doi:10.1080/10408340500526766, https://doi.org/10.1080/10408340500526766. Dostupné z: https://doi.org/10.1080/10408340500526766

[37] Puzis, R.; Elovici, Y.; Dolev, S.: Fast algorithm for successive computation of group betweenness centrality. *Physical Review E*, ročník 76, Nov 2007: str. 056709, doi:10.1103/PhysRevE.76.056709. Dostupné z: https://link.aps.org/doi/10.1103/PhysRevE.76.056709

[38] Zhao, J.; Wang, P.; Lui, J. C.; aj.: I/O-efficient calculation of H-group closeness centrality over disk-resident graphs. *Information Sciences*, ročník 400-401, 2017: s. 105–128, ISSN 0020-0255, doi:https://doi.org/10.1016/j.ins.2017.03.017. Dostupné z: https://www.sciencedirect.com/science/article/pii/S0020025517305960

Příloha A

Obsah přiložených souborů

- `data` ..... obsahuje použité datasety
  - `email_eucore` ..... dataset Eu-core
    - `email_features.csv` ..... rozdělení do skupin pro Eu-core
    - `email_nodes.csv` ..... hrany datasetu Eu-core
  - `facebook_large` ..... dataset Facebook
    - `musea_facebook_target.csv` ... rozdělení do skupin pro Facebook
    - `musea_facebook_edges.csv` ..... hrany datasetu Facebook
  - `git` ..... dataset Git
    - `musae_git_target.csv` ..... rozdělení do skupin pro Git
    - `musea_git_edges.csv` ..... hrany datasetu Git
  - `lastfm_asia` ..... dataset LastFM
    - `lastfm_asia_target.csv` ..... rozdělení do skupin pro LastFM
    - `lastfm_asia_edges.csv` ..... hrany datasetu LastFM
  - `twitch_gamers` ..... dataset Twitch
    - `large_twitch_features.csv` ..... rozdělení do skupin pro Twitch
    - `large_twitch_edges.csv` ..... hrany datasetu Twitch
- `prakticka_cast_bp_turekmat.ipynb` ..... zdrojové kódy implementace
- `thesis_latex` ..... zdrojová forma práce ve formátu LaTex
- `thesis.pdf` ..... bakalářská práce ve formátu PDF