Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra informatiky a výpočetní techniky

Bakalářská práce

Analýza sociální sítě přátel

Plzeň 2013 Marek Naggy

Originální zadání

Místo této stránky bude vloženo originální zadání BP.

Prohlášení

Prohlašuji, že jsem bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů.

V Plzni dne 2. května 2013

Marek Naggy

Poděkování

Na tomto místě bych rád poděkoval Ing. Michalu Nyklovi za ochotné vedení práce, cenné rady a čas, který mi věnoval při konzultacích.

Abstract

Analysis of social network of friends

This thesis introduces the opportunities that social network analysis offers by analyzing online large-scale social network of friendship. Thus, you will find here a design of the web crawler, which has successfully crawled more than a million web pages located on the online social network called Lidé.cz. In the thesis you will also find statistics which show socio-demographics of this server. It is also stated what kind of information you can obtain through the analysis of online social network content. The above mentioned social network is then verified against the theoretical models. Furthermore, there are rankings which contain the most important users of this network. These rankings were created by using basic Centrality measures methods (such as Degree, Closeness and Betweenness centrality). In addition the small-world theory is verified.

Key words: social network analysis, Centrality measures, web crawler, small-world theory, scale free networks, Lidé.cz

Analýza sociální sítě přátel

Práce se pomocí analýzy rozsáhlé online sociální sítě přátelství zaměřuje na přiblížení možností, které analýza sociálních sítí nabízí. Naleznete zde návrh webového robota, který prošel více než milion webových stránek nacházejících se na online sociální síti Lidé.cz. Jsou zde uvedeny statistiky zabývající se sociodemografií tohoto serveru a je zde uvedeno, jaké informace pomocí analýzy obsahu sociálních sítí lze získat. Získaná síť je poté verifikována oproti teoretickým modelům. Též jsou zde pomocí několika základních metod Centrality measures (Degree, Closeness a Betweenness centrality) určeny nejvýznamnější uživatelé nacházející se v této síti. Práce se též zabývá ověřením hypotézy malého světa.

Klíčová slova: analýza sociálních sítí, Centrality measures, webový robot, teorie malého světa, bezškálové sítě, Lidé.cz

Obsah

1 Úvod **1**

2 Sociální sítě a jejich analýza **2**
2.1 Poloměr, hustota a další pojmy z teorie grafů 3
2.2 Modely a vlastnosti komplexních sítí 5
2.2.1 Bezškálové sítě a mocninný zákon 6
2.3 Small-world experiment 7
2.4 Centrality measures 8
2.4.1 Degree centrality 8
2.4.2 Closeness centrality 9
2.4.3 Betweenness centrality 9

3 Online sociální síť Lidé.cz **11**
3.1 Smluvní podmínky a autorský zákon 11
3.2 Předpokládaná velikost 11
3.3 Omezení stahování 12
3.4 Které webové stránky procházet? 13

4 Získání dat **15**
4.1 Výběr vhodného webového robota 15
4.2 Tvorba webového robota 16
4.2.1 Ukládání dat a návrh databázové struktury 16
4.2.2 Architektura, algoritmy a použité technologie 17
4.2.3 Implementační poznámky 19
4.3 Získaná data 21
4.4 Známé problémy 22

5 Statistické výstupy a jejich porovnání s dostupnými daty **24**
5.1 Jaké informace zjišťovat a z jakého důvodu? 24
5.2 Obyvatelstvo 25
5.3 Cizí jazyky a vzdělání 27
5.4 Mladiství 28
5.5 Oblíbené sporty a hudba 30
5.6 Shrnutí kapitoly 30

6 Analýza získané sítě **31**
6.1 Výběr vhodného frameworku 31
6.2 Základní informace o síti 32
6.3 Výsledky Centrality measures a jejich porovnání 35

7 Závěr **39**
7.1 Budoucí práce 40

A Uživatelské příručky **47**
A.1 Webový robot SNABot 47
A.2 Framework SNAP 48
A.3 SNABot-nastroje 48

B Formáty souborů **50**

C Rozšíření statistik **51**

D Porovnání výsledků aproximace s přesným výpočtem **52**

E Výsledky Centrality measures pro orientovanou síť a aproximace **54**

F Jádro sítě **57**

1 Úvod

S rozvojem online sociálních sítí (online social network – OSN), odehrávající se v posledních letech, se otvírá nová možnost, jak zkoumat strukturu lidské společnosti a jejich vztahů. Touto možností je analýza sociálních sítí (social network analysis – SNA). Tento analytický nástroj je používán např. v marketingu, sociologii či kriminalistice. Práce se zabývá analýzou sociální sítě přátel, která byla získána z OSN Lidé.cz. Obecně se však OSN nemusíme omezovat. Tyto metody lze uplatnit v mnoha dalších sociálních sítích od spolupráce mezi herci, telefonními hovory až po kriminální síť.

Cílem práce není komplexní analýza zvolené OSN (nenaleznete zde např. analýzu komunit či dynamiky), ale přiblížit čtenáři možnosti, které SNA nabízí a ukázat, proč jsou její metody využívány např. FBI [25] či armádními zpravodajci [26].

Po prvních dvou kapitolách, věnujících se zejména pojmům z teorie grafů, představení a analýze zvolené OSN, následuje kapitola zabývající se automatickým sběrem dat pro další analýzu. Vzhledem k tomu, že pro naši analýzu byla zvolena OSN, je v této kapitole popsán postup návrhu webového robota, který umožní její automatické získání.

Pátá kapitola si klade za cíl motivovat např. zaměstnance marketingových či reklamních agentur a ukázat, jaké informace lze z dat uchovaných v OSN získat. Výsledky jsou průběžně porovnávány s podobnými průzkumy, jež jsou prováděny „klasickou“ dotazníkovou formou. Tento krok se snaží určit, do jaké míry lze považovat data, která uživatelé uvádějí, za spolehlivá.

Poslední kapitola se zabývá samotnou analýzou získané sítě. Naleznete zde některé základní vlastnosti této sítě a porovnání, na kolik odpovídá teoretickým modelům, které jsou uvedeny v druhé kapitole. Další částí této kapitoly jsou žebříčky sestavené dle metod, které umožňují v síti identifikovat klíčové osoby, diskuse získaných výsledků a jejich porovnání s podobnými pracemi.

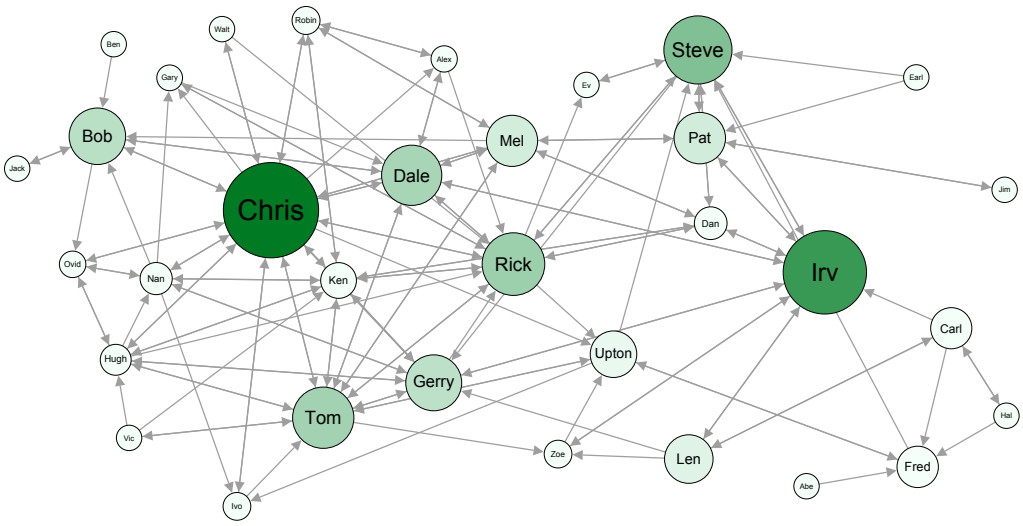
V získané síti se též snažím ověřit teorii malého světa, založenou S. Milgramem roku 1967 [29]. Ta inspirovala divadelní hru a následně televizní film, který tuto hypotézu proslavil tvrzením: „Každý na této planetě je oddělen jen šesti jinými lidmi. Šest kroků od sebe. Mezi námi a kýmkoli jiným na této planetě. Prezident Spojených států. Benátský gondoliér... “ [10].

Práce je inspirována výzkumem A.L. Barabásiho [5, 10] zkoumajícím různé typy komplexních sítí a představující jejich tzv. bezškálový model. Dále pracemi, které zkoumají jiné OSN (např. Facebook.com, Flickr.com aj.) [18, 30, 6]. Za české jmenujme [21] analyzující OSN CouchSurfing.com či [24] zkoumající sociální síť organizátorů zážitkových akcí.

2 Sociální sítě a jejich analýza

Pojem *sociální síť* pochází ze sociologie a je formulován jako skupina sociálních vztahů nebo vazeb mezi aktéry, které vzájemně propojují [14, 23]. Tyto vazby mohou být kladné či záporné, mít stejnou nebo rozdílnou váhu a orientaci.

Sociální síť lze reprezentovat grafem (viz část 2.1) nebo maticí sousednosti [23, 26]. Graf, který reprezentuje sociální síť se nazývá sociogramem [14, 23]. V sociogramu jsou aktéři zobrazeni jako vrcholy (body) a jejich vztahy jako hrany (vazby, které vrcholy spojují). Jeho ukázku můžete vidět na obr. 2.1 (pro vizualizaci byla použita data z [12] a program Gephi [11]). V tomto obr. je velikost a barva vrcholu dána jeho „oblibou“ mezi ostatními vrcholy (jedná se o orientovanou vazbu – „považuji ho za přítele“). Výpočet je založen na metrice Betweenness centrality (viz kapitola 2.4.3).



Obr. 2.1: Ukázka sociogramu.

Online sociální síť (Online social network – OSN) je v [17] definována jako webová služba, která uživatelům umožňuje vytvořit si vlastní profil[1], navazovat s dalšími uživateli spojení a prohlížet si jejich profily nebo jejich spojení. OSN se tak snaží přenést části reálných sociálních sítí do virtuálního prostředí.

[1]Vlastní webovou stránku, která obsahuje informace o uživateli.

Za první OSN je považována SixDegrees.com, spuštěná v roce 1997 a ukončena v roce 2000 [17]. Mezi dnešní nejznámější OSN patří Facebook[2], Twitter[3] nebo Google+[4]. Za české jmenujme Lidé.cz[5] nebo Spolužáci.cz[6].

Analýzou sociální sítě (jak reálné, tak virtuální) rozumíme zkoumání, detekování a interpretování vzorů, které tvoří sociální vazby mezi aktéry [16]. Oblasti využití této analýzy jsou široké, od využití v marketingu, modelování přenosů nemocí, zkoumání možného napojení na kriminální organizace, až po analýzu politického hlasování [10, 26, 27].

2.1 Poloměr, hustota a další pojmy z teorie grafů

V této části naleznete několik pojmů z teorie grafů, které budou v textu dále používány a jejich aplikaci na sociální síť, která bude dále zkoumána. Jako zdroje byly použity [4, 18, 23].

Graf, jeho orientace a váhy hran

Jako *graf* G označujeme dvojici $G = (V, E)$, kde V je konečná neprázdná množina vrcholů a E je konečná množina hran, které mezi sebou propojují dvojice vrcholů z množiny V. Počet vrcholů V budeme značit jako n a počet hran E jako m.

Pokud budou vrcholy spojeny hranou, bude v textu dále značeno, že vrchol v_i má kontakt na vrchol v_j a hrana bude značena jako $\{v_i, v_j\}$. V našem případě, budou vrcholy představovat uživatele OSN a hrany jejich virtuální přátelství.

Jako *orientovaný graf* uvažujeme takový graf, u kterého rozlišujeme směr odkud (z výchozího vrcholu) kam (do cílového vrcholu) hrana vede tzn. hrana $\{x, y\} \neq \{y, x\}$. U *neorientovaného grafu* tento směr neuvažujeme tzn. hrana $\{x, y\} = \{y, x\}$.

Graf označíme jako *nevážený*, pokud jeho hrany nejsou ohodnoceny váhou[7], případně pokud platí $\omega(e) = 1; \forall e \in E$, kde ω je funkcí určující váhu hrany. *Váženým grafem* naopak označíme graf, u kterého rozlišujeme váhy jednotlivých hran.

[2] `http://www.facebook.com`
[3] `http://www.twitter.com`
[4] `http://plus.google.com`
[5] `http://www.lide.cz`
[6] `http://www.spoluzaci.cz`
[7] Váha může představovat např. cenu, vzdálenost, kapacitu apod.

Stupeň vrcholu

Jako *stupeň vrcholu* v $d_G(v)$ označujeme počet hran z množiny E, pro které platí $\{v, x\}$ či $\{x, v\}$. Pokud mluvíme o neorientovaném grafu, nezáleží, zda bude uzel v cílový nebo počáteční. U orientovaného grafu označujeme jako $d_G^+(v)$ počet výstupních hran z vrcholu v. Pro počet vstupních hran používáme značení $d_G^-(v)$. V případě váženého grafu počítáme nejen s počtem hran, ale i s jejich váhou.

V našem případě stupeň vrcholu v vyjadřuje, kolik má uživatel v virtuálních přátel (výstupních hran) a kolik uživatelů ho za přítele považuje (vstupní hrany). V části 2.4.1 naleznete metriku, která pracuje výhradně s touto informací.

Cesta a nejkratší cesta

Cestou v grafu G označujeme libovolnou posloupnost přechodů mezi vrcholy v_a a v_z za pomoci hran $\{v_a, v_b\}$,$\{v_b, v_c\}$ ... $\{v_y, v_z\}$, ve které se každý vrchol v_i objevuje pouze jednou.

Nejkratší cestou mezi dvěma vrcholy rozumíme cestu, jejíž ohodnocení je ze všech existujících cest minimální. Pokud graf obsahuje více komponent (K_i, K_j viz dále), délka nejkratší cesty z vrcholu $v_a \in K_i$ do vrcholu $v_z \in K_j$ je nekonečná.

V našem případě cesta znamená posloupnost uživatelů, která je zapotřebí pro komunikaci mezi uživateli v_a a v_z. Tuto komunikaci je možné zprostředkovat pomocí přátel.

Komponenta

Komponentou K_i v grafu G označíme podgraf (podmnožina vrcholů a hran původního grafu), neexistuje-li cesta z vrcholu $v_i \in K_i$ do vrcholu $v_j \in K_j$, přičemž platí $K_i \cap K_j = \emptyset$. Jinými slovy – žádný vrchol z K_i není spojen hranou s vrcholem patřící do K_j. Pokud z každého vrcholu v existuje konečná cesta do všech vrcholů, má graf jednu komponentu.

Pokud by v grafu bylo více komponent, znamená to, že uživatelé z komponenty K_i, nemohou za pomoci svých kontaktů, komunikovat s uživateli z komponenty K_j.

Hustota grafu

Hustota grafu je dána jako poměr hran existujících a hran všech možných. Pro neorientovaný graf tedy $\Delta = \frac{2m}{n(n-1)}$, pro graf orientovaný pak $\Delta = \frac{m}{n(n-1)}$. Kde m je počet hran a n počet vrcholů. Její hodnota se pohybuje v intervalu $< 0, 1 >$.

V reálných rozsáhlých sítích tato hodnota většinou nabývá malých hodnot [16].

Poloměr grafu

Poloměr grafu D je roven ohodnocení nejdelší z nejkratších cest všech dvojic vrcholů grafu G. V případě, že je graf rozdělený na více komponent, platí $D = \infty$. V tom případě má smysl uvažovat poloměr největší komponenty.

Průměrná délka nejkratší cesty

Průměrná délka nejkratší cesty (Average shortest path length – ASPL) l v grafu G, je definována jako aritmetický průměr délek nejkratších cest mezi všemi možnými dvojicemi vrcholů v grafu G (viz (1) a [15]).

$$l = \frac{1}{n(n-1)} \sum_{i \neq j} g(v_i, v_j) \tag{1}$$

Kde n je počet vrcholů a $g(v_i, v_j)$ je délka nejkratší cesty mezi vrcholy v_i a v_j. Pokud graf obsahuje více komponent, určujeme ASPL v největší komponentě.

Tuto hodnotu lze odhadnout jako $l \sim \ln(n)/\ln\ln(n)$.

Jak v lidské společnosti (ve smyslu propojení pomocí přátel), tak v dalších sítích, jako např. spolupráce herců v Holywoodu, citačních sítí aj. (viz [10]) je tato hodnota relativně malá – ASPL mezi občany USA je uváděna jako 6,5 (viz kapitola 2.3).

2.2 Modely a vlastnosti komplexních sítí

Text v této části vychází ze zdrojů [10, 15, 5]. Pro modelování komplexních sítí se v současnosti používají tři hlavní paradigmata: teorie náhodných grafů, Watts-Strogatzův model a model bezškálových sítí.

Model náhodných sítí (1959) [10]. předpokládá, že každá dvojice vrcholů je spojena hranou s pravděpodobností p. Pokud bychom tedy uvažovali $p = 0$, žádné vrcholy spojeny nebudou, naopak, pokud $p = 1$ budou spolu spojeny všechny vrcholy. Takto vytvořené sítě, mají relativně malý poloměr, ale jejich další vlastnosti neodpovídají empirickým výsledkům získaným z reálných sítí[8]. Zejména kvůli absenci vrcholů s několikanásobně větším stupněm než je stupeň průměrný, které se v reálných sítích běžně vyskytují. Dále kvůli malému koeficientu shlukování, ve kterém se odráží neschopnost podobných vrcholů vytvářet tzv. shluky, ve kterých je většina vrcholů vzájemně propojena.

[8]Ty můžete nalézt v [5], jedná se např. o síť webových stránek, spolupráce herců aj.

Model, který problém s malým koeficientem shlukování odstraňuje, se nazývá *Watt-Strogatzův model* (1998) [10]. Ten používá vrcholy uspořádané do kruhu, jenž jsou propojeny se svými K sousedy, kde K určuje počet propojených sousedů ($K/2$ na každé straně). Poté jsou některé vrcholy propojeny jako v modelu předchozím – tento krok vytváří tzv. slabé vazby[9] zmenšující poloměr sítě. Tento model však stále nevyřešil existenci vrcholů s vysokým stupněm. Ten řeší až *model bezškálových sítí* (2001) [5].

2.2.1 Bezškálové sítě a mocninný zákon

Při zkoumání reálných sítí, ať již rozvoden elektřiny, spolupráce herců v Holywoodu nebo zkoumání struktury internetu, se ukázalo, že některé vrcholy mají oproti jiným několikanásobně větší stupeň. Takovéto vrcholy budeme dále označovat jako *centra*.

V předchozích modelech je nepřítomnost center dána Poissonovým rozdělením pravděpodobnosti, kterým se u vrcholů řídí rozdělení jejich stupně. Pro vysvětlení vzniku center je třeba vzít v úvahu, že vrcholy jsou do sítě přidávány postupně a neexistují v síti od jejího počátku, jak uvažovaly modely předchozí. Starší vrcholy tak mají více času na získání více kontaktů. Dle [10] tento samotný fakt pro vznik center nestačí. Dalším pravidlem pro připojování vrcholů je tzv. *preferenční připojování*. To říká, že pravděpodobnost připojení nového vrcholu k vrcholům stávajícím je úměrná jejich stupni a „atraktivitě“ pro nové vrcholy. Čím větší stupeň tedy vrchol má a čím je pro nové vrcholy „atraktivnější“, tím je pravděpodobnější, že získá další kontakty.

Na obr. 2.2 můžete vidět vývoj bezškálové sítě, kde každý obrázek představuje přidání nového vrcholu (znázorněný prázdným kroužkem), přičemž se každý nový vrchol může připojit právě ke dvěma vrcholům. Nové vrcholy se řídí preferenčním připojováním (pouze podle stupně) a jak je z obrázku vidět, vznikají tak centra s vysokým stupněm.

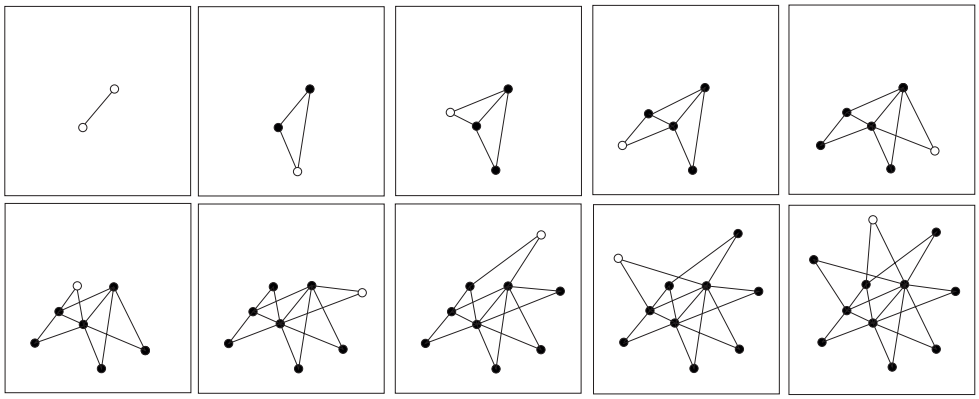
Tyto dva faktory odstraňují výše zmiňované Poissnovo rozdělení a nahrazují ho tzv. *mocninným zákonem*. Tento zákon se řídí vzorcem (2),

$$N(k) = k^{-\gamma} \tag{2}$$

kde $N(k)$ značí četnost výskytu vrcholů stupně k. Jako γ je označen exponent konektivity. Tento exponent má v reálných sítích nejčastěji hodnotu $2 \leq \gamma \leq 3$ [10].

Sítě, řídící se mocninným zákonem mají několik společných vlastností. Díky centrům mají relativně malý poloměr a jsou odolné proti náhodným výpadkům. V sítích, ve kterých platí $\gamma \leq 3$, pak lze náhodně odebrat téměř všechny vrcholy a síť se stále

[9] V [10, 14] se uvádí, že právě slabé vazby hrají v sociální interakci důležitou roli např. při hledání zaměstnání nebo při rozšiřování fám.



Obr. 2.2: Vývoj bezškálové sítě – obrázek byl převzat z [10].

nerozpadne na více komponent [10]. Na druhou stranu, pro bezškálové sítě platí, že pokud bychom vrcholy odebírali cíleně, stačí jich odebrat pouze několik a síť se rozpadne – v tomto případě bychom volili právě centra.

2.3 Small-world experiment

Roku 1967 provedl S. Milgram experiment [29], který měl ověřit kolik prostředníků je zapotřebí pro spojení dvou náhodných lidí v USA. Při tomto experimentu rozeslal více než sto složek různým lidem v Nebrasce, kteří měli za úkol přeposlat tyto složky cílovému člověku v Massachusetts[10] a dodržet pravidlo, že je složku možné přeposlat pouze osobě, kterou osobně znají. Cílem bylo zjistit, na kolik je lidská společnost propojená (formou známých a přátel).

Výsledky ukázaly, že tyto složky dorazily k cílovému člověku za použití dvou až deseti prostředníků, přičemž jejich průměrný počet byl 5,5. Je nutné podotknout, že výsledek může být poněkud zavádějící. Délka těchto cest byla dána nejlepším odhadem lidí, kteří topologii této sítě neznali. Na zkreslení výsledku se taktéž mohlo projevit to, že z původních 160 složek jich k cíli dorazilo pouze 27,5%, tj. jen 44.

Po průzkumu řetězců Milgram nalezl několik dalších zajímavostí. Cílové osobě dorazilo 48% složek přes cesty, kde se jako poslední zprostředkovatelé vyskytovali pouze tři různé osoby. To může naznačovat, že některé cesty jsou využívány více než ostatní. U podobného experimentu (označovaného jako Kansaský) zjistil, že jak ženy, tak muži přeposílali složky stejnému pohlaví častěji – přibližně o 78%.

Tento experiment inspiroval divadelní hru s názvem *Šest stupňů odloučení*, která byla později zfilmována. V té se mylně uvádí, že fenomén šesti kroků platí pro celý svět [10]. Poslední studie, viz [6], zkoumající více než 720 miliónů uživatelů OSN

[10]Vzdálený z výchozího bodu přibližně 2 300 km.

Facebook.com z několika zemí však ukazuje, že tato hodnota je ještě menší. Jako ASPL uvádí 4,74 – tedy „pouze“[11] 3,74 stupně odloučení.

2.4 Centrality measures

Centrality measures (CM) jsou kolekcí metod, umožňujících měření tzv. *centrality* jednotlivých vrcholů, která určuje jejich významnost v síti [18, 23, 26]. Vrcholy s vysokou centralitou mají v síti výhodnější pozici, např. ve smyslu možnosti kontroly toku informací nebo možnosti ovlivňovat vrcholy ostatní.

Mezi tři hlavní metody, které umožňují centralitu měřit, patří Degree centrality (C_D), Closeness centrality (C_C) a Betweenness centrality (C_B) [23, 26]. V dalším textu budou popsány základní přístupy k těmto metodám, nejčastěji označované jako Freemanovy[12]. Pro každou z těchto metod existuje několik dalších variant (např. Bonachiho přístup pro C_D, Eigenvector vycházející z C_C aj.), které tyto základní přístupy rozšiřují (viz [23]). Výsledky všech metod lze normalizovat, což umožňuje vzájemné porovnání mezi podobnými sítěmi.

2.4.1 Degree centrality

Degree centrality je nejjednodušší metoda, kdy je centralita určena stupněm uzlu[13] (viz část 2.1). U orientované sítě se může jednat o stupeň výstupní, vstupní nebo kombinaci obou.

Vrchol s větším stupněm je zvýhodněn tím, že má více možností, jak v síti komunikovat. Tím se stává méně závislým na vrcholech dalších. Zároveň může jiným vrcholům sloužit jako prostředník při komunikaci a z této služby těžit. V orientované síti jsou vrcholy s vysokým vstupním stupněm označovány jako prominentní nebo prestižní, vrcholy s vysokým výstupním stupněm pak jako vrcholy vlivné [23].

Normalizovaná verze této metriky viz vzorec (3) a [18],

$$C_D(v_i) = \frac{d(v_i)}{n-1} \tag{3}$$

kde $d(v_i)$ je stupeň vrcholu v_i a n celkový počet vrcholů v síti.

[11] Zde je možné se na problém dívat ze dvou stran. Z jedné strany se skutečně jedná o malé číslo. Na druhou stranu se jedná o několik okruhů známých, což může být „velká psychologická vzdálenost“ [29].

[12] Jejich autorem je sociolog Linton C. Freeman, viz [20].

[13] U vážené sítě můžeme použít tzv. Weighed degree centrality, která pro výpočet používá nejen počet hran, ale i jejich váhu.

Nevýhodou této metody je, že vrcholy s vysokou hodnotou C_D nemusí být ty s největší centralitu nebo mohou být centrální pouze lokálně [23]. Naopak výhodou této metody je nízká výpočetní složitost.

2.4.2 Closeness centrality

Closeness centrality – přesnější metoda, která analyzuje, jak je vrchol blízko k dalším vrcholům. Čím je blíže, tím efektivněji dokáže s dalšími vrcholy komunikovat. Taktéž tím vzrůstá šance, že komunikace bude úspěšná (zpráva nebude jinými uzly zamítnuta apod.). Další výhodou vrcholu s vysokou hodnotou C_C je, že se jej pravděpodobně budou ostatní vrcholy snažit využít při komunikaci, čímž tento vrchol získává na důležitosti [23].

Výpočet této metriky je dán jako převrácená hodnota součtu délek nejkratších cest sítě. Normalizovaná verze viz vzorec (4) a [18],

$$C_C(v_i) = \left[\frac{1}{n-1} \sum_{j \neq i}^{n} g(v_i, v_j) \right]^{-1} \tag{4}$$

kde v_i je výchozí vrchol, v_j vrchol cílový, n počet vrcholů v síti a $g(v_i, v_j)$ je délka nejkratší cesty z vrcholu v_i do vrcholu v_j.

Tato metoda narozdíl od C_D dokáže odhalit i vrcholy menšího stupně, které jsou však na základě své „polohy“ v síti významné [23, 26]. Nevýhodou je její výpočetní složitost, která je pro vážené sítě $O(nm + n^3)$ a pro nevážené $O(nm + n^2)$, viz [8].

2.4.3 Betweenness centrality

Betweenness centrality je komplexní metoda, která analyzuje, jak často vrchol leží na nejkratší cestě při komunikaci mezi vrcholy dalšími [18, 23].

Vrcholy s vysokou hodnotou C_B jsou důležité, protože např. umožňují efektivní komunikaci mezi různými komunitami. Svého postavení tak mohou využívat a vybírat „servisní poplatky“, rozhodnout se blokovat nežádoucí zprávy nebo izolovat některé vrcholy. Tímto mohou do značné míry kontrolovat dění v síti [23, 26]. Výpočet C_B se pro vrchol v_i vypočítá jako:

$$C_B(v_i) = \sum_{v_s \neq v_i \neq v_t} \frac{\sigma_{st}(v_i)}{\sigma_{st}} \tag{5}$$

kde σ_{st} je počet nejkratších cest mezi vrcholy v_s a v_t (pokud mají stejné ohodnocení, může jich být více) a $\sigma_{st}(v_i)$ je počet nejkratších cest mezi vrcholy v_s a v_t, které procházejí vrcholem v_i [18].

Normalizace této metriky se provede vydělením $C_B(v_i)$ s $(n-1)(n-2)$ pro síť orientovanou a $\frac{(n-1)(n-2)}{2}$ pro síť neorientovanou [32]. Kde n je počet vrcholů v síti.

Výhodou C_B je nalezení vrcholů, které jsou v síti nejvíce využívány při předávání zpráv, za předpokladu využívání nejkratších cest. Její nevýhodou je její výpočetní složitost, která je uváděna jako $O(mn + n^2 \log n)$ pro vážené sítě, pro nevážené pak $O(mn)$, viz [8]. To platí při použití algoritmu uváděného v [13] jinak je pro vážené i nevážené $O(n^3)$.

3 Online sociální síť Lidé.cz

Sociální síť *Lidé.cz* je největší českou OSN[1], kterou provozuje firma *Seznam.cz, a.s.* V době svého vzniku (1997) sloužila k vyhledávání e-mailových adres[2]. Později se transformovala do dnešní podoby, umožňující uživatelům mezi sebou komunikovat, seznamovat se, vzájemně si prohlížet profily apod.

3.1 Smluvní podmínky a autorský zákon

Pátý bod třetí části smluvních podmínek Lidé.cz (viz [36]) říká, že uvedením osobních informací o své osobě, uživatel souhlasí s jejich zveřejněním předem neomezenému okruhu osob. Z toho plyne, že při zpracování informací získaných z této OSN by neměl být porušen *zákon o ochraně osobních údajů*.

Autorský zákon takovéto vytěžování databáze – zejména pro komerční účely – zakazuje. Zároveň ho však dle §92 povoluje a) pro osobní potřebu, b) *pro účely vědecké nebo vyučovací*, c) pro účely veřejné bezpečnosti nebo správního či soudního řízení [1].

3.2 Předpokládaná velikost

Oficiální informace o aktuální velikosti této OSN nejsou uváděny. Počet online uživatelů se pohybuje v rozmezí 20 - 30 tisíc. Tisková zpráva z konce března 2007, zveřejněná prostřednictvím serveru Lupa.cz[3], se zmiňuje o překonání hranice milionu uživatelů s vlastním profilem.

Společnost Netmonitor ve veřejných výstupech z prosince 2012 uvádí, že přibližný počet unikátních měsíčních přístupů (RU_M) na subdoméně `http://profil.lide.cz` byl 610 tisíc. Zde je pro srovnání několik údajů z předešlých období (hodnoty uvedeny v počtu RU_M):

Červen 2007 – 1,1 milionu; Leden 2010 – 1,5 milionu (maximum v obr. 3.1); Prosinec 2011 – 830 tisíc; Květen 2012 – 710 tisíc.

[1] Dle společnosti NetMonitor (`http://www.netmonitor.cz`).
[2] `http://onas.seznam.cz/cz/o-firme/historie-firmy/1997`
[3] `http://www.lupa.cz`

Obr. 3.1 zobrazuje relativní objem vyhledávání této OSN pomocí vyhledávače Google.com. Z obrázku je vidět postupný útlum, který je pravděpodobně způsoben vstupem zahraničních OSN (např. Facebook.com) na český trh.

Obr. 3.1: Relativní objem vyhledávání OSN Lidé.cz dle aplikace *Google Trends*[4].

Jelikož průměrný denně strávený čas návštěvníka na této OSN je relativně vysoký (cca 35 min), předpokládám, že se jedná o „stálé návštěvníky“. Současný počet aktivních profilů[5] se tak dle mého názoru blíží průměrnému počtu RU_M za poslední rok. Tedy 600 až 800 tisícům. Přičemž z obr. 3.1 usuzuji, že tento počet stagnuje či pozvolna klesá.

3.3 Omezení stahování

Po prozkoumání podmínek používání [35] a smluvních ujednání [36] Lidé.cz zjistíme, že se provozovatel o automatickém procházení nezmiňuje (oproti tomu např. při automatickém procházení OSN Facebook.com je nutné žádat o povolení[6]). Soubor `robots.txt`[7] na adrese `http://lide.cz/robots.txt`, obsahuje pouze řádky:

```
User-Agent: *
Disallow:
```

Ty robotům povolují následovat veškeré odkazy [2].

Automatické procházení není omezeno ani HTML metatagem:

```
<meta name="robots"content="noindex, nofollow">
```

Stále je však možné, že si provozovatel nebude přát, aby tato OSN byla procházena neznámými roboty (např. kvůli omezení zátěže serveru) a bude se proti tomu bránit.

[4] `http://www.google.com/trends`
[5] Neaktivním uživatelům, kteří se více než 6 měsíců nepřihlásí, je jejich účet smazán [36].
[6] `http://www.facebook.com/apps/site_scraping_tos.php`
[7] Soubor, který povoluje robotům procházet určité části webové prezentace.

Jako příklad uveďme blokaci IP adres s vysokým počtem požadavků, nebo použití CAPTCHA[8] kódů. V takovém případě by pro naše účely bylo vhodné zvážit změnu OSN.

Další možností by mohlo být získání dat z webových archivů, jako např. `http://www.archive.org`. Ty ale nejsou příliš aktuální a obsahují pouze zlomek požadovaných webových stránek. Alternativní možností by bylo získat data z webové cache Google.com, což je místo, kde Google.com uchovává kopie webových stránek, které prošel jejich robot [28]. Do této cache se dá dostat např. přes adresu:

`http://webcache.googleusercontent.com/search?q=cache:adresa`

Případně ve vyhledávači Google.com vyhledat `cache:adresa`. Tato cache však též neobsahuje veškeré požadované stránky a počet přístupů do ní může být omezen.

3.4 Které webové stránky procházet?

Každý uživatel, který si na OSN Lidé.cz založí účet, dostane přidělený profil, na unikátní adrese `http://profil.lide.cz/JménoUživatele/profil`. Tato úvodní stránka obsahuje stručné informace o uživateli, několik jeho přátel a nástěnku (sloužící jako vzkazník, kde spolu uživatelé mohou veřejně komunikovat). Dalšími stránkami jsou:

- `o-mne`
- `moji-pratele`
- `maji-me-v-pratelich`
- `fotogalerie`
- `videa`

Stránka `o-mne` obsahuje veškeré informace, které o sobě uživatel uvedl, jako např. jeho záliby, věk, bydliště aj.

Stránka `moji přátelé` je rozdělena na dvě části. První část `moji-pratele` obsahuje seznam uživatelů, které majitel profilu považuje za své přátele. Druhá část, `maji-me-v-pratelich` je seznam uživatelů, kteří považují majitele profilu za svého přítele. Tyto dva seznamy se při nepovinné autorizaci (žádosti o navázání přátelství není třeba potvrzovat), na rozdíl od některých jiných OSN, nemusí shodovat. V tom případě se jedná o orientovaná spojení.

[8] Completely Automated Public Turing test to Tell Computers and Humans Apart.

Pokud vezmeme v úvahu výše uvedené, je vhodné získávat data pouze ze stránek `o-mne` a moji přátelé. Z časových důvodů – předpokládám stažení až 800 tisíc profilů (viz část 3.2) – jsem se rozhodl stahovat pouze údaje ze stránek `moji-pratele` a `o-mne`. Z dat, získaných ze stránek `moji-pratele`, bude vytvořena a dále analyzována síť přátelství uživatelů této OSN. Data ze stránek `o-mne` budou použita pro statistické účely, jejichž výsledky budou porovnány s dostupnými údaji.

4 Získání dat

V této kapitole naleznete algoritmy a požadavky na webového robota (dále jen robota), několik poznámek k implementaci a způsob, jakým budou data ukládána. Na konci kapitoly jsou uvedeny známé problémy a několik statistik, které jsem při průchodu OSN shromáždil. Uživatelskou příručku pro práci s roboty naleznete v příloze A.1.

4.1 Výběr vhodného webového robota

Při výběru vhodného robota jsem vycházel z několika již implementovaných řešení (Web-Harvest 2.0[1], JSpider 0.5[2], JoBo 1.4[3], crawler4j[4] aj.). Základní podmínkou výběru byla obstojná dokumentace a splnění co největšího počtu těchto, mnou požadovaných, vlastností:

- Možnost ukládat data do databáze, CSV či XML souboru.
- Možnost přerušit stahování a možnost pokračovat tam, kde robot naposledy skončil.
- Logování akcí robota.
- Generování základních statistik stahování.
- Nastavení robota vlastním zdrojovým kódem či pomocí XML souboru.
- Přítomnost DOM[5] parseru, nebo jiné možnosti výběru části HTML stránky, která bude ukládána.
- Nastavení doby prodlevy mezi stahováním jednotlivých stránek.
- Možnost zpracování nevalidního HTML.

Všechny tyto požadavky splňuje Scrapy [3]. Webový robot, napsaný v Pythonu, kterého je možné pomocí vlastního zdrojového kódu libovolně upravit.

[1] `http://web-harvest.sourceforge.net`
[2] `http://j-spider.sourceforge.net`
[3] `http://www.matuschek.net/jobo-menu`
[4] `http://code.google.com/p/crawler4j`
[5] Document Object Model – umožňuje přistupovat k datům uloženým v XML nebo HTML souborech pomocí stromové struktury, viz `http://www.w3.org/DOM`.

4.2 Tvorba webového robota

Jako základ, pro tvorbu webového robota, jsem použil framework Scrapy (v. 0.14) [3]. Framework je možné používat i na slabších strojích (při běhu využívá cca 50 MB paměti).

4.2.1 Ukládání dat a návrh databázové struktury

Pro další práci, jsem se rozhodl ukládat data v *relační databázi* z následujících důvodů: a) poskytuje dostatečnou rychlost; b) oproti XML souboru bude objem dat menší; c) data jsou ukládána postupně a nehrozí tak jejich ztráta; d) potřebná data lze snadno a rychle vyhledávat; e) oproti grafovým databázím mám s jejich používáním zkušenosti. To umožní rychlejší vývoj a zbude tak více času na získání a analýzu dat.

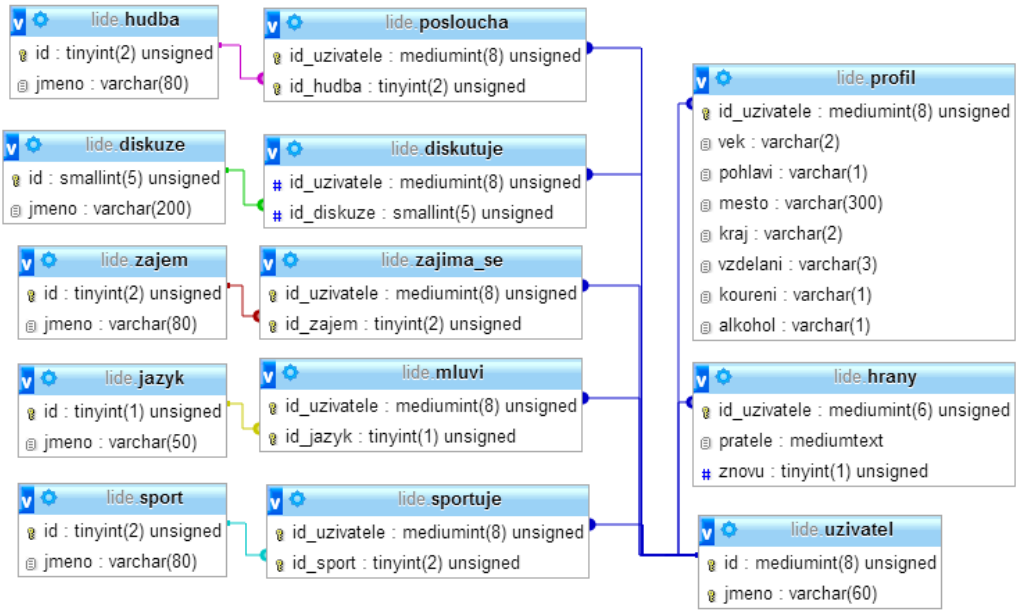
Jako cílovou relační databázi jsem vybral MySQL[6] Community Server (v. 5.5.25a), provozovanou na lokálním stroji. Toto řešení nabízí dostatečnou výkonnost, je šířeno pod GPL[7] licencí, je přenositelné a s využitím nástroje phpMyAdmin[8] nabízí příjemné uživatelské rozhraní. Robot byl na databázi napojen pomocí knihovny MySQLdb[9].

Na obr. 4.1 můžete vidět strukturu navržené databáze. Mezi tabulkami `profil`, `hrany` a `uzivatel` jsem zvolil kardinalitu 1:1. To, v případě potřeby omezení datového objemu, umožní pracovat pouze s částí databáze. V případě odpojení se též omezí počet „poloprázdných“ záznamů, které by vznikly při použití jedné tabulky.

U tabulky `hrany` jsem pro atribut `pratele` zvolil kompaktnější způsob uchovávání dat. Tento atribut je uchováván jako textový seznam. Hlavním důvodem pro použití této konstrukce, byl předpoklad relativně velkého průměrného počtu přátel jednotlivých uživatelů – dle vizuální kontroly cca 30 - 50. Testy ukázaly, že tento způsob uložení může ušetřit při předpokládaném počtu 40 milionů hran až 400 MB.

Stejný formát pak lze použít i pro reprezentaci sítě v programech, které s ní dále pracují (jako např. Pajek XXL, Gephi aj., viz část 6.1).

[6] `http://www.mysql.com`
[7] GNU General Public License – licence pro svobodný software (viz `http://www.gnu.org`).
[8] `http://www.phpmyadmin.net`
[9] `http://sourceforge.net/projects/mysql-python`



Obr. 4.1: Struktura použité databáze.

4.2.2 Architektura, algoritmy a použité technologie

Minimalizace rizika odpojení

Při návrhu robota jsem vycházel z potenciálního hrozícího odpojení robota od OSN. V první řadě jsem se proto rozhodl pro získání sítě přátelství, která bude dále analyzována (viz kapitola 6). Data pro statistické účely pak stahovat paralelně z webové cache společnosti Google.com (viz část 3.3), či je stahovat, po získání sítě přátelství, přímo z OSN Lidé.cz. Takto rozložené zatížení by mělo stahování urychlit a zmenšit riziko odpojení.

Jako prodleva mezi stahováním jednotlivých stránek byla zvolena jedna vteřina. To by mělo zvýšit pravděpodobnost, že robot nebude odpojen, případně bude odpojen až po získání většího množství dat.

Robota je možné v libovolnou chvíli pozastavit a při dalším spuštění pokračovat tam, kde naposledy přestal. Tato vlastnost by také měla zmenšit riziko odpojení.

Získání potřebných dat

Jelikož veškeré profily, mají stejnou strukturu (liší se pouze vzhledem[10]) a jedná se o relativně malý HTML dokument, je pro získání dat vhodné použít DOM parser

[10]Uživatelé si mohou vybrat z několika tzv. skinů, které pomocí různých kaskádových stylů mění vzhled jejich profilu.

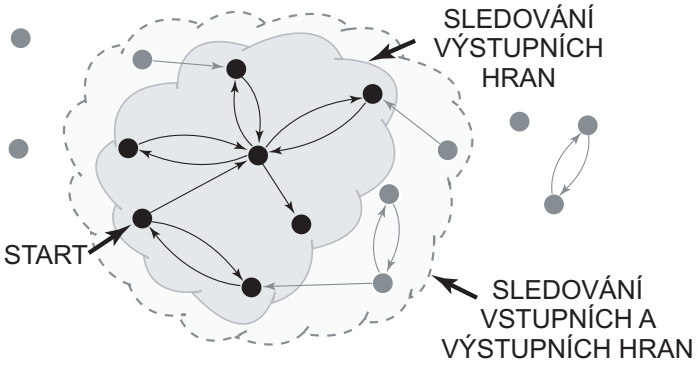
a potřebná data získat pomocí *XPath*[11] cest. Při změně struktury získávaných dokumentů je však nutné aktualizovat zastaralou XPath cestu. Protože zdrojový kód zpracovávaných HTML souborů je tvořen více než tisícem řádků, využil jsem pro určení XPath cest nástrojů jako XPath Helper či FireBug.

Strategie procházení

Jako výchozí body, ze kterých robot začne OSN procházet, jsem určil pseudonáhodné uživatele s přihlédnutím na geografickou rozmanitost jejich bydliště a počtu jejich přátel.

Pro procházení OSN jsem použil algoritmus *DFS*[12] (ve frameworku Scrapy realizovaný zásobníkem), který by, na rozdíl od *BFS*[13], měl i při odpojení poskytnout přesnější údaje např. o exponentu konektivity nebo o průměrné délce nejkratších cest (ASPL), viz [30]. Další algoritmus, pomocí kterého lze OSN procházet, můžete nalézt v [18]. Jedná se o tzv. *Uniform Sampling*. Ten je založen na procházení náhodných adres profilů, což umožňuje nalezení profilů z více komponent. Vzhledem k předem neomezenému počtu variací, tvořících uživatelské jméno, ale není pro naše účely vhodný.

Jak bylo dříve zmíněno, z časových důvodů jsem procházel pouze stránku `mojipratele` (obsahující výstupní hrany). To mohlo způsobit nekompletní zpracování komponenty, viz obr. 4.2. Z obrázku je vidět, že při tomto přístupu se mohou vyskytovat uživatelé, patřící do stejné komponenty, kteří nebudou nalezeni, pokud na ně nesměřuje žádná vstupní hrana. Lze však předpokládat, že pro zkoumanou síť nebudou příliš významní.



Obr. 4.2: Ukázka nalezených vrcholů při následování výstupních hran v orientované síti. Převzato z [30].

[11] Viz `http://www.w3schools.com/xpath`.
[12] Depth-first search – prohledávání do hloubky.
[13] Breadth-first search – prohledávání do šířky.

Rozdělení programů a popis algoritmů

Pro procházení OSN Lidé.cz vznikli celkem tři roboti. SNABot – pro procházení sítě přátelství. SNABot-Profily – pro procházení profilů a SNABot-Centra – pro procházení center (viz popis problému s centry v části 4.4).

Algoritmus nejdůležitějších akcí robota SNABot:

1) Projdi výchozí stránku/y a ulož do zásobníku žádané odkazy[14].

2) Vyjmi ze zásobníku odkaz a odešli požadavek na server. Pokud je zásobník prázdný a nejsou naplánovány další požadavky, přejdi na 7).

3) Ze získané odpovědi získej seznam přátel.

4) Ulož do databáze aktuálního uživatele a jeho přátele (tabulka `uzivatele`) a výstupní hrany uživatele (tabulka `hrany`). Pokud bude přátel více než 500, nastav záznamu v `hrany.znovu` příznak na 1.

5) Přidej do zásobníku odkazy na uživatelovi přátele.

6) Pokud je zásobník prázdný a neexistují další požadavky, přejdi na 7), jinak opakuj 2).

7) Ukonči program.

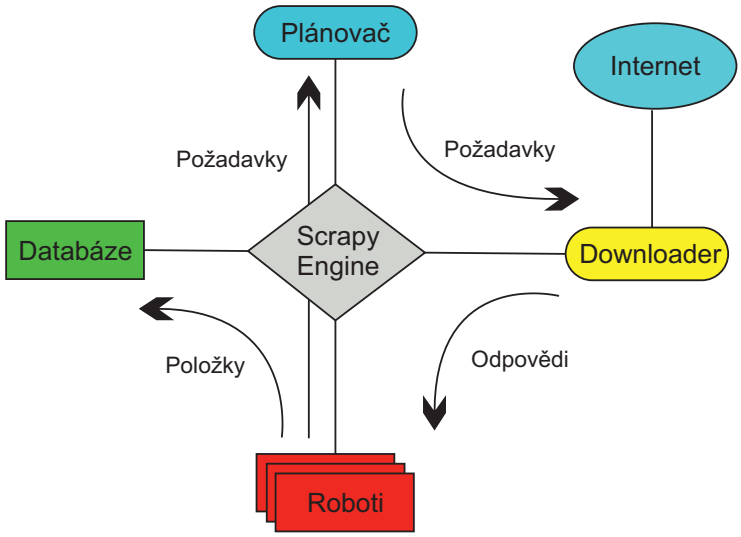
Další roboti pracují obdobně. Robot *zpracovávající profily*, prochází iteračně databázi a v každé iteraci naplánuje několik odkazů, které budou dále procházeny. Robot pro *extrakci kontaktů center*, řešící problém s centry (viz část 4.4 – problém s centry), prochází iteračně záznamy, které byly označeny ve čtvrtém kroku robotem SNABot. Prochází však mobilní verzi této OSN umístěnou na `http://m.lide.cz`[15], generuje a zpracovává předem známé odkazy (získané z uživatelských jmen). Tím výše uvedený problém řeší. Architekturu robotů, založených na frameworku Scrapy, můžete vidět na obr. 4.3.

4.2.3 Implementační poznámky

V této části je stručně uvedeno použití některých specifických konstrukcí frameworku Scrapy. Další implementační podrobnosti naleznete v oficiální dokumentaci projektu, viz [3].

[14] Odkazy, na které lze aplikovat pravidla dána regulárními výrazy, viz část 4.2.3

[15] Na té existuje stránkování, které umožňuje seznamem přátel listovat. Je ovšem omezeno pouze stranami 1 až 100. Tzn. maximálně 1200 přáteli. Toto stránkování však lze „oklamat“ pomocí editace URL adresy. V té je možné nastavit offset a pomocí něj projít veškeré přátele.



Obr. 4.3: Ukázka architektury webových robotů, založených na frameworku Scrapy. Obrázek byl převzat z [3].

Robotovi je možné nastavit pravidla, které přiřazují stránkám různé akce a lze s jejich pomocí obsah stránek zpracovávat. Toho lze docílit kódem:

```
rules = (
       Rule(SgmlLinkExtractor(allow = "profil.lide.cz/(.+)/moji-pratele/"),
            callback ="pratele_parse"),
       Rule(SgmlLinkExtractor(allow = "profil.lide.cz/(.+)/o-mne/"),
            callback = "profil_parse")
)
```

Kde řetězec přiřazený `allow` je regulární výraz. Pokud odkaz bude odpovídat alespoň jednomu regulárnímu výrazu, uplatní se na něj metoda mající stejný název, jako řetězec přiřazený proměnné `Callback`. Pokud robot nalezne odkazy, na které žádná pravidla uplatnit nelze, budou ignorovány.

Výchozí adresy se určují jako `start_urls = ["www.prvni.cz", ... ]`. Na výchozí adresy nejsou uplatňována pravidla – tzn. výchozí adresy nebudou zpracovány! Proto doporučuji výchozí adresy např. „podstrčit“ v lokálním souboru jako `start_urls = ["file:start.html"]`. Adresy obsažené v souboru již budou standardně zpracovány.

V získaném HTML dokumentu lze pomocí XPath cest vybírat data následovně:

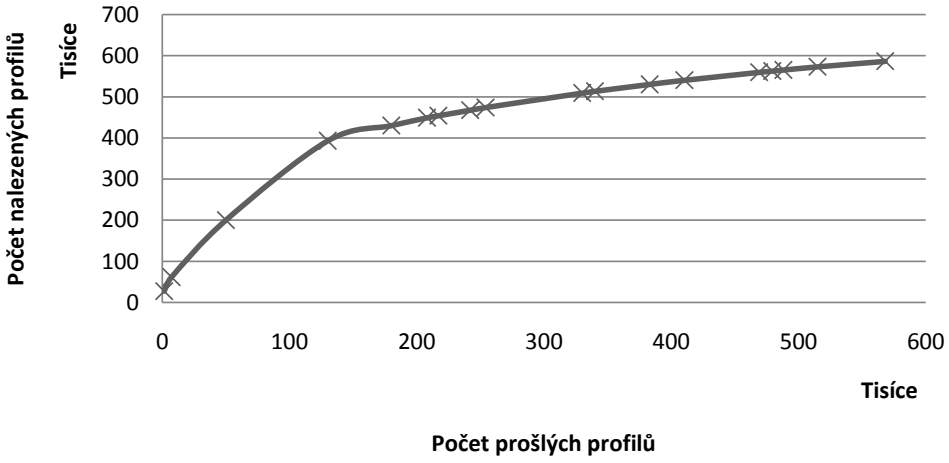
```
xpath = HtmlXPathSelector(response)
xpath.select("//div[@id='profileTitle']/h2/text()").extract()[0]
```

Kde `response` je získaný HTML dokument a řetězec v metodě `select` XPath cesta, pomocí níž lze v dokumentu vybrat požadovaná data. Metoda `extract` vrací unicódový řetězec vybraný pomocí XPath cesty.

Časovou prodlevu mezi stahováním jednotlivých stránek lze měnit pomocí proměnné `download_delay`. V aktuálním nastavení roboti zapisují veškeré informace do souborů `log-rrrr-mm-dd.txt`. Stupeň výpisů lze zmírnit např. pouze na chybové hlášky změnou hodnoty `LOG_LEVEL` na `ERROR` v souboru `settings.py`. Další změnu nastavení je možné provést v témže souboru.

4.3 Získaná data

Sběr dat pro vytvoření sítě přátelství (tzn. stránek `moji-pratele`) neprobíhal kontinuálně, robot byl několikrát pozastaven (nejčastěji po 12 hodinách běhu). Sběr dat trval přibližně měsíc, přičemž *robot nebyl od OSN odpojen*. Při průchodu OSN Lidé.cz bylo celkem nalezeno 589 374 uživatelů (včetně duplicit, viz část 4.4). Z obr. 4.4 můžete vidět, že pro nalezený počet uživatelů téměř platí Paretovo pravidlo [10] (pravidlo 80/20), v následujícím kontextu: „Po průchodu a zpracování prvních 20% uživatelů, nalezneme 80% uživatelů z celkového počtu“.



Obr. 4.4: Počet nalezených uživatelů.

Celkový počet úspěšně prošlých stránek `moji-pratele` je 574 988. Po odečtení duplicitních uživatelů (cca 10 tisíc), zjistíme, že za měsíc bylo smazáno přibližně 4,5 tisíce profilů.

Při stahování profilů z webové cache společnosti Google.com byl robot odpojen (viz 4.4 – Odpojení od cache Google.com). Mimo to se ukázalo, že tato cache obsahuje pouze cca 58% požadovaných stránek. Data ze stránek `o-mne` proto byla získána přímo z cílové OSN, po získání sítě přátelství. Stahování trvalo cca 40 dní, přičemž

opět neprobíhalo nepřetržitě. Další zajímavostí je, že při posledních iteracích stahování – cca o tři měsíce později po zahájení stahování stránek `moji-pratele` – byl poměr neexistujících stránek (smazaných profilů) k počtu nalezených profilů 1,6 : 23,5. Tzn., že *za tři měsíce bylo smazáno cca 7% uživatelských účtů.* To, dle mého názoru, může být jednak důsledkem vysoké dynamiky OSN nebo to může ukazovat na postupný „odliv“ uživatelů (viz část 3.2).

Robot procházející mobilní verzi OSN, prošel za cca 2,5 hodiny všechny označené záznamy tvořené 58 centry s více než 500 přáteli. Ze 7 tisíc získaných stránek tak mohl získat až 27 tisíc nových hran[16].

Roboti prošli více než 1,1 milionu webových stránek. Průměrný počet stažených stránek za minutu se pohyboval v intervalu 50-55 stránek/min. Při stahování bylo odesláno na server cca 380 MB požadavků a staženo cca 4,75 GB dat. Z těchto dat vznikla databáze mající přibližně 9,6 milionu řádků o velikosti 370 MB. SQL Dump této databáze naleznete na přiloženém DVD.

4.4 Známé problémy

Při získávání dat se vyskytlo několik problémů, které budou dále popsány.

Problém s centry

Ukázalo se, že počet zobrazovaných přátel na stránce `moji-pratele` je omezen horní hranicí 500 přátel (řazeno abecedně). Tento problém se podařilo odstranit dodatečným robotem, který centra znovu projde (viz část 5.2.2 – robot pro extrakci kontaktů center).

Problém s diakritikou v názvech profilů

Při stahování dat bylo zjištěno, že existuje minoritní množství uživatelů, kteří si v minulosti vytvořili profil se jménem, obsahující diakritiku. To v současnosti není možné. Tyto profily jsou pak spojeny s profily bez diakritiky (např. uživatel = uzivatel). Tvar zobrazovaného jména v seznamech přátel není normalizován. Tato skutečnost vytváří v databázi duplicitní záznamy, mezi které jsou vstupní hrany rozděleny. Před samotnou analýzou OSN je nutné tyto duplicity odstranit.

Tento problém řeší aplikace SNABot-Nastroje, která umožňuje profily sjednotit a přesměrovat chybné vstupní hrany duplicit. Pro budoucí práci doporučuji robota upravit tak, aby zpracovával pouze účty bez diakritiky. Nebude tak třeba účty sjednocovat a dle naměřených údajů, se projde cca o 10 tisíc stránek méně.

[16]Každé centrum mělo 500 hran, tzn. 29 tisíc jich již existovalo. Z každé stránky, vyjma posledních, bylo získáno 8 hran.

Odpojení od cache Google.com

Dalším problémem se ukázalo paralelní stahování dat z cache Google.com. Již při testování byl robot odpojen přibližně po zaslání sta požadavků. Nepomohlo ani zvýšení časové prodlevy mezi stahováním jednotlivých stránek. V [22] autor, který získával data z Google Scholar[17], uvádí cituji: „Experimentálně bylo zjištěno, že ideální je pozastavit program na jednu hodinu po každých 150 uskutečněných dotazech. Je ovšem možné, že po větším počtu přístupů z jedné IP adresy, bude nutné limity zpřísnit.“ Kvůli neefektivitě (150 získaných stránek za hodinu oproti 3600), jsem od paralelního stahování dat z této cache upustil a rozhodl se toto řešení využít pouze v případě, pokud bude robot SNABot z OSN Lidé.cz odpojen.

Nepřetržitá změna dat

Posledním problémem, který byl předem znám, ale nebyl očekáván v takovém rozsahu, je neustálá změna OSN. Jejich velikost a neustálá změna znemožňuje získání všech požadovaných informací ve stejném čase. V našem případě to např. způsobuje existenci uživatelů, které někdo považuje za přátele, ale než se robot dostal k jejich zpracování, jejich účet byl smazán. Nebylo tak možné získat jejich přátele a profily.

[17] `http://scholar.google.com`

5 Statistické výstupy a jejich porovnání s dostupnými daty

V této kapitole naleznete několik statistik, které jsou následně porovnány s daty Českého statistického úřadu (ČSÚ) či jinými zdroji. Tyto informace lze použít např. pro určení největší cílové skupiny. Pokud se potvrdí, že většina dat souhlasí s již provedenými měřeními, mohou být přínosná i pro sociology (viz dále). Následující informace byly získány z databáze profilů, obsahující více než 550 tisíc profilů, kterou se podařilo vytvořit robotem SNABot-Profily (viz kapitola 4).

Tvorba grafů a výsledků je z části zautomatizována. Aplikace SNABot-Nastroje umožňuje vytvořit CSV soubory (viz příloha A.3) s požadovanými daty a z těch pak pomocí R[1] skriptů (nacházejících se na přiloženém DVD) vygenerovat dále uvedené grafy. Při aktualizaci dat je tak možné následující statistiky poměrně snadno aktualizovat.

5.1 Jaké informace zjišťovat a z jakého důvodu?

Cílem této části je vytvořit ukázkovou množinu otázek, týkající se dat získaných z cílové OSN, které budou dále zkoumány. Na základě výsledků pak získat představu, nakolik jsou data udávaná uživateli shodná s údaji uváděnými jinými zdroji a zjistit tak jejich relevantnost.

V případě jejich shody by stálo za zvážení, zda neprovádět některé kvantitativní sociologické průzkumy analýzou dat získaných ze sociálních sítí. Tato alternativa by oproti dotazníkové formě poskytla větší počet respondentů a ušetřila by náklady spojené s distribucí a následným ručním vyhodnocováním dotazníků.

Jako ukázkovou množinu jsem zvolil následující otázky:

Populace zkoumané OSN

- Jaké je věkové složení uživatelů?
- Jak jsou uživatelé rozmístěni dle jednotlivých krajů?
- Platí Milgramem pozorovaný úkaz z části 2.3 – stejné pohlaví se kontaktuje (v našem případě přátelí) častěji?

[1] Jedná se o software umožnující, pomocí jazyka R, statistické výpočty a tvorbu grafů. Viz `http://www.r-project.org`.

Cizí jazyky a vzdělání

- Jaké je složení uživatelů dle nejvyššího dosaženého vzdělání?
- Kolika cizími jazyky uživatelé hovoří?
- Bude nejvíce uváděným cizím jazykem angličtina[2] následována němčinou?
- Souvisí počet cizích jazyků, kterými uživatel hovoří, s dosaženým vzděláním?

Mladiství

- Kolik procent mladistvých kouří a/nebo konzumuje alkohol?
- Kouří a konzumují alkohol více chlapci, nebo dívky?

Oblíbené sporty a hudba

- Jaké jsou nejpopulárnější sporty a hudební žánry?

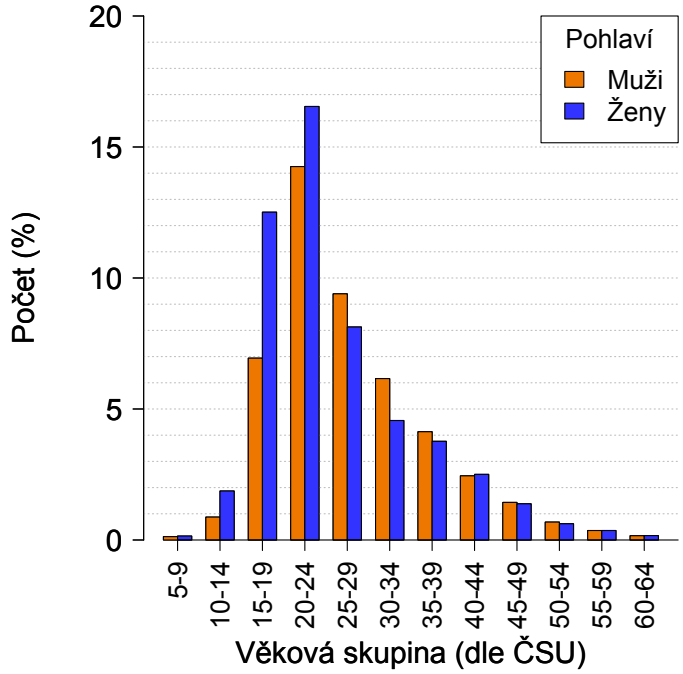
5.2 Obyvatelstvo

Z obr. 5.1 je vyplývá, že nejvíce zastoupenou skupinou v této OSN jsou lidé ve věku 15 - 29 let. Průměrný věk je 26 let. 55,8% uživatelů svůj věk neuvádí. V populaci ve věku 15 - 19 jasně převládají ženy, další věkové skupiny jsou poměrně vyrovnané. Celkové zastoupení uživatelů tvoří 45,5% mužů a 54,5% žen.

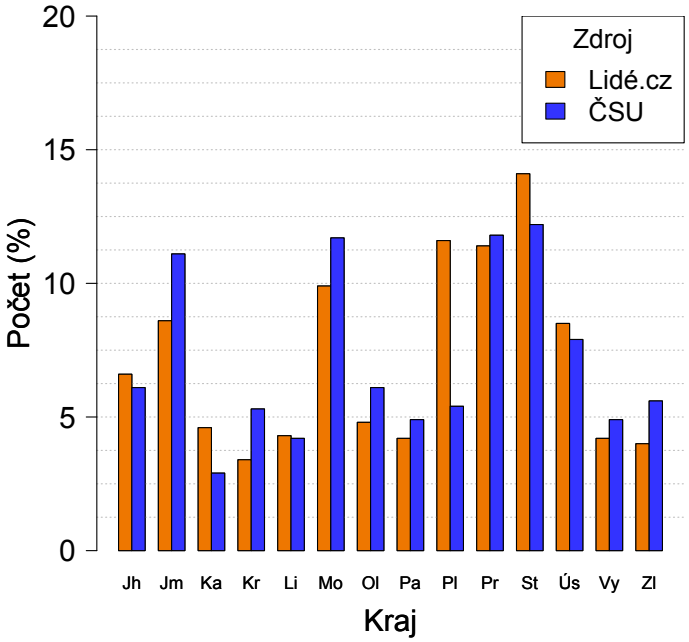
Věkové složení uživatelů ani složení pohlaví neodpovídají údajům uváděnými ČSÚ (50,9% žen a 49,1% mužů, viz soubor `csu-kr-obyvatele.xls` na přiloženém DVD). Neshodují se ani s oficiálními údaji uváděných na OSN[3] a to sice: 53% mužů a 47% žen. Ty vycházejí z měření společnosti Netmonitor.cz (jako vzorek bylo použito 2 889 respondentů). Je však třeba vzít v úvahu odlišnou metodologii těchto měření. Oficiální statistiky vychází z údajů poskytnutých návštěvníky této OSN nikoliv ze složení profilů.

Z obr. 5.2 vyplývá, že složení uživatelů této OSN procentuálně odpovídá složení obyvatelstva v ČR dle krajů (s tolerancí 2 - 3% u jednotlivých krajů). Výraznou anomálii tvoří pouze uživatelé z Plzeňského kraje. V tomto kraji je zkoumaná OSN velmi populární, zejména v Plzni. 72,3% uživatelů z Plzeňského kraje uvádí ve svém profilu město Plzeň. Data z ČSÚ uvádí, že pouze 56,4% obyvatel z Plzeňského kraje se nachází v okresech Plzeň-město, Plzeň-jih a Plzeň-sever (či pouze 32,3% Plzeň-město). 19,4% uživatelů kraj neuvedlo.

[2] Slovenština zde není zahrnuta, jelikož na zkoumané OSN ji není možné vybrat.
[3] Viz `http://onas.seznam.cz/cz/lide-cz.html`.



Obr. 5.1: Věkové složení uživatelů OSN Lidé.cz dle pětiletých věkových skupin a pohlaví.



Obr. 5.2: Porovnané složení uživatelů OSN Lidé.cz s obyvateli ČR dle krajů.

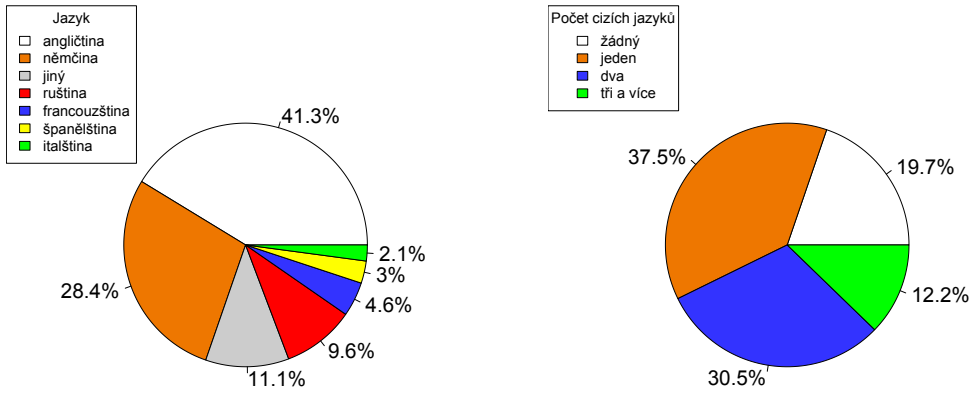
Úkaz, pozorovaný Milgramem, o častějším využívání kontaktů stejného pohlaví se neprokázal. Naopak pouze 12,5% ze všech přátelství tvoří přátelství mezi muži a 30,1% mezi ženami.

5.3 Cizí jazyky a vzdělání

Informace o nejvyšším dosaženém vzdělání byly zjišťovány u uživatelů ve věku 15 let a starších. Nejvíce je v této OSN zastoupena skupina uživatelů s maturitou – 34,7%, středním vzděláním bez maturity – 33,5% a základním vzděláním – 23,7%. Vysokoškolské vzdělání je zde zastoupeno cca 8%. Své vzdělání uvádí 59,4% uživatelů.

ČSÚ uvádí tyto údaje[4]: základní vzdělání 18%, střední vzdělání bez maturity 33%, střední vzdělání s maturitou 34% a vysokoškolské vzdělání 15%. Z toho vyplývá, že tuto OSN využívá větší procentuální část lidí se základním vzděláním, nežli je průměr v ČR. Naopak zde nejsou příliš zastoupeni lidé vysokoškolsky vzdělaní.

Počet cizích jazyků, kterými uživatelé údajně hovoří, naleznete na obr. 5.3b. Procentuální zastoupení jednotlivých cizích jazyků na obr. 5.3a. Tyto výsledky jsou podobné těm, které jsou uváděny v [33] (jedná se o průzkum mezi občany ve věku 18–69 let, zkoumáno bylo 9 500 domácností). V tomto výzkumu uvádí 40% respondentů znalost jednoho jazyka, 24% dvou a 7% tří a více. 30% dotazovaných nemluví žádným cizím jazykem.



(a) Poměr výskytu jednotlivých cizích jazyků.

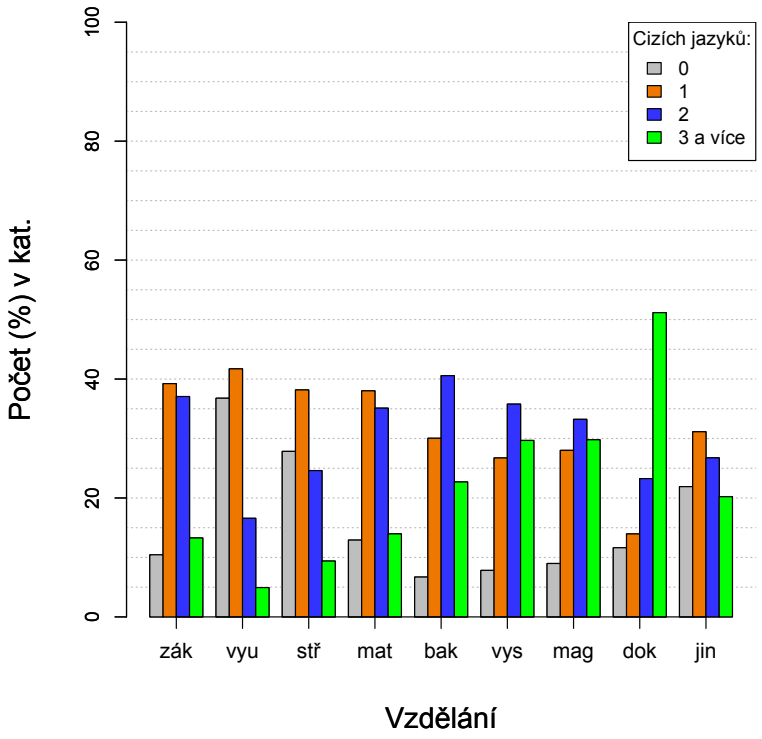
(b) Poměr počtu cizích jazyků, kolika údajně uživatelé hovoří.

Obr. 5.3: Znalosti cizích jazyků.

[4]Viz prezentace O. Nývlta na přiloženém DVD.

Nejčastěji uváděnými cizími jazyky (dle [33]) jsou: angličtina – 45%, němčina – 26% a ruština – 23%. Výzkum [34] (z roku 2002, kde bylo dotázáno 900 respondentů) uvádí, že ruština je zastoupena v největší míře u starších generací. Vzhledem k malému zastoupení, které ve zkoumané OSN tvoří, to bude pravděpodobně důvod, proč není uváděna častěji.

Na obr. 5.4 naleznete grafické znázornění závislosti počtu jazyků, které uživatelé této OSN uvádějí, na jejich dosaženém vzdělání. [34] uvádí, že nejčastější forma výuky cizích jazyků (u lidí ve věku 15 - 30 let, bylo možné vybrat více možností) je: výuka na SŠ 60,5%, ZŠ 42% a VŠ 21%. Za zajímavý fakt tak považuji to, že lidé se základním vzděláním uvádějí znalost více jazyků nežli vyučení lidé a velmi podobnou znalost lidem se středním vzděláním s maturitou. U vysokoškolsky vzdělaných lidí je výrazný nárůst znalosti tří a více jazyků. U těchto skupin tvoří téměř 25% a více.



Obr. 5.4: Závislost znalosti počtu cizích jazyků na dosaženém vzděláním.

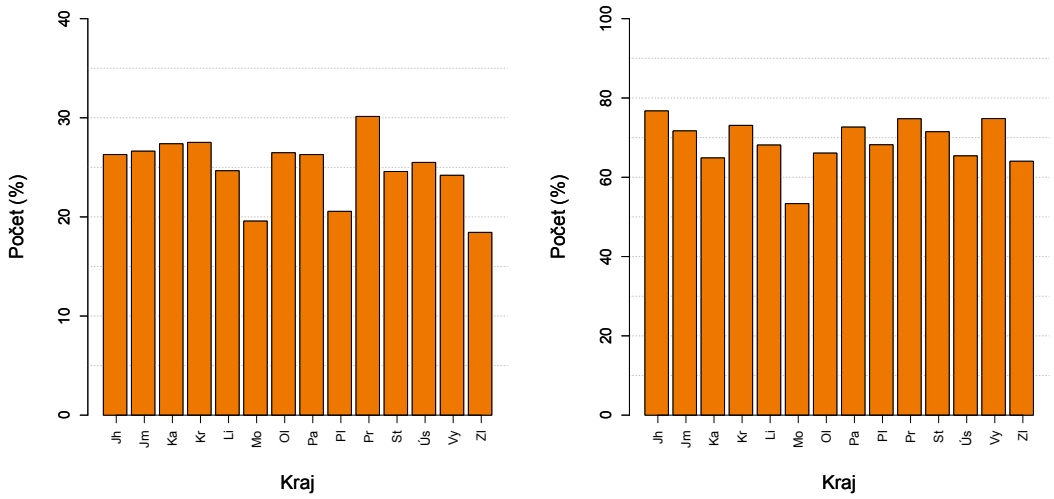
5.4 Mladiství

V síti existuje 27 048 uživatelů mladších 18 let (tj. 11% z uživatelů uvádějících svůj věk). Z toho jich 43,3% uvádí, že alespoň příležitostně kouří (cca 17,7%) nebo konzumují alkohol (cca 41,1%). Pouze 16,8% uživatelů mladších 18 let uvádí, že

nekouří ani nekonzumují alkohol. 25% jich o sobě tyto údaje neuvádí a 17,8% uvádí pouze jeden údaj.

V následujícím textu jsou vynechány osoby, které o sobě neuvedly obě informace (tj. 42,8% mladistvých). Nejvíce (23,9%) je zastoupena skupina mladistvých konzumujících alkohol, která uvedla, že nekouří. Naopak kuřáků nekonzumujících alkohol je velmi málo (1,4%). Kuřáků konzumujících alkohol je 16,5%. Zdá se tak, že většina mladistvých má prvotní zkušenost s alkoholem a část z nich poté začne být i kuřáky. Podrobnější údaje naleznete v příloze C.

Porovnání mladistvých kuřáků (obr. 5.5a) a konzumentů alkoholu (obr. 5.5b) dle krajů naleznete na obr. 5.5 (vynecháni mladiství bez vyplněných informací). Z těchto grafů vyplývá, že nejvíce mladistvých kuřáků je v Praze, nejméně pak v Plzeňském, Zlínském a Moravskoslezském kraji. Nejméně mladistvých konzumentů alkoholu je opět v Moravskoslezském kraji.



(a) Porovnání mladistvých kuřáku dle jednotlivých krajů.

(b) Porovnání mladistvých konzumentů alkoholu dle jednotlivých krajů.

Obr. 5.5: Porovnání mladistvých kuřáků a konzumentů alkoholu dle krajů.

V [31] (studie omezená na 16 leté studenty) je uvedeno, že mezi mladistvými se vyskytuje 25,7% kuřáků (27,1% chlapců a 24,2% dívek) a přibližně 79% mladistvých, kteří v posledních 30 dnech pily alkohol. Průzkum byl proveden v roce 2011 a týkal se 3 913 studentů. V případě omezení na mladistvé ve věku 16 let, jich ve zkoumané OSN 74,4 % uvádí konzumaci alkoholu a 26,5%, že jsou kuřáci (vzorek tvořily cca 4 tisíce uživatelů a nebyli bráni v úvahu mladiství bez vyplněné informace).

Dále se ze získaných dat podařilo zjistit (viz tab. 5.1, která bere v úvahu mladistvé ve věku 16 let uvádějící oba údaje), že mezi kuřáky je větší procentuální zastoupení dívek (34,4% oproti 27,8% chlapců). Konzumace alkoholu je poměrně vyrovnaná (78,8% dívek oproti 76,9% chlapců).

Tab. 5.1: Kouření a konzumace alkoholu dle pohlaví.

Pohlaví	Celkem	Alkohol	Kouření
Chlapci	1430	1127	398
Dívky	2536	1951	872

5.5 Oblíbené sporty a hudba

Nejčastěji uváděný hudební žánr je rock. Ten uvádí 40,8% uživatelů. Dále pop – 34,8% a hip-hop – 26,3%. Nejméně uváděnými žánry jsou opera a opereta 1,4% a swing 1,7%. Práce [19] (bylo dotazováno 440 respondentů) na prvních dvou místech uvádí též rock a pop (neuvažujeme-li soundtrack).

Mezi nejčastěji uváděné sporty v OSN Lidé.cz patří kolektivní hry s míčem – 33,5%, cyklistika – 29,3% a zimní sporty – 25,6%. Naopak nejméně uváděnými sporty jsou parasporty (parašutismus apod.) – 1,7% a jogging – 2%. Zajímavou infografiku o oblíbenosti sportů uvádí [37]. Zde se na prvním místě umístil fotbal, na druhém cyklistika a na třetím hokej. Jako nejčastěji provozovaný sport se uvádí cyklistika.

5.6 Shrnutí kapitoly

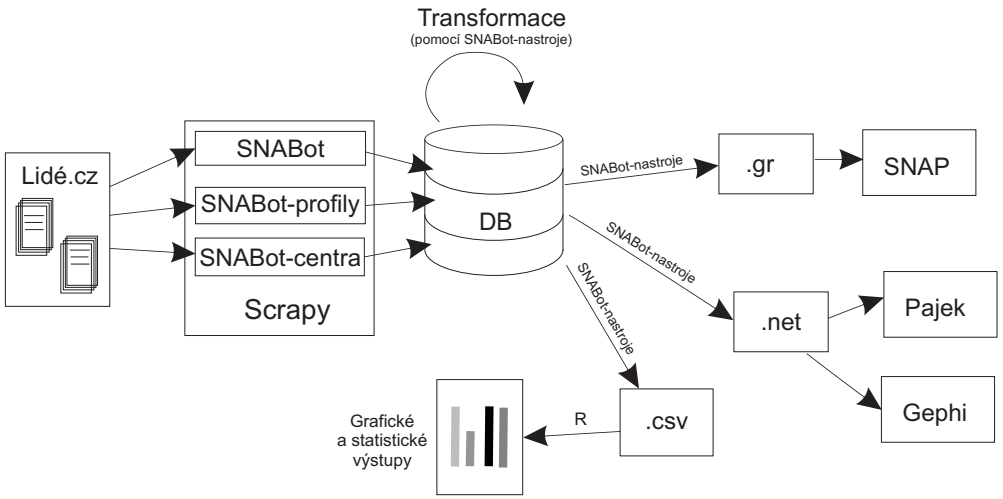
V této kapitole bylo uvedeno několik statistik, které byly porovnány s výsledky z jiných zdrojů. Většina z nich odpovídala výsledkům z jiných zdrojů s poměrně malou tolerancí.

Lze tak doporučit používat analýzu obsahu OSN jako alternativní formu k získávání kvantitativních informací. Její hlavní výhodou je snadné shromáždění velkého objemu dat, které, jak bylo v této kapitole naznačeno, jsou dostatečně relevantní. Tuto alternativu je však nutné používat s rozmyslem a uvědomit si, že některá data mohou být, zejména proto, že si uživatelé uvědomují jejich dostupnost, částečně zkreslená.

Údaje uvedené v této kapitole lze např. využít k určení, jak velkou cílovou skupinu může oslovit reklamní sdělení. Např. reklama s výbavou pro rybáře nebo s běžeckými potřebami nenalezne příliš potencionálních zájemců. Naopak reklama např. s výbavou cyklistickou, lyžařkou či se vstupenkami na koncert oblíbeného rockového zpěváka by jich mohla oslovit až několik desítek, ne-li stovek, tisíc.

6 Analýza získané sítě

V této kapitole naleznete informace o získané síti, ověření, zda odpovídá bezškálové topologii (viz 2.2.1), a žebříčky uživatelů sestavené dle metod Centrality measures (CM), které jsou následně porovnány. Počátek kapitoly tvoří stručné představení použitých programů a frameworků. Obr. 6.1 zobrazuje proces, jakým byla data získána a zpracována.



Obr. 6.1: Proces použitý pro získání informací z cílové OSN.

6.1 Výběr vhodného frameworku

Částečný seznam nástrojů, umožňujících analýzu sítí, naleznete např. v [7]. Některým z nich se však získanou síť, kvůli její velikosti, nepodařilo ani načíst. U dalších, vzhledem k neparalelní implementaci a kvadratické či vyšší asymptotické složitosti, nebylo možné hodnoty centralit (C_B a C_C) v reálném čase vypočítat. Pomocí programu Pajek XXL jsem zjistil, že přesný neparalelní výpočet (jedné z centralit) by na stroji s procesorem AMD Phentom II X4 955 3,20 GHz trval přibližně měsíc.

Small-world Network Analysis and Partitioning (SNAP)

Pro výpočet centralit jsem zvolil framework Small-world Network Analysis and Partitioning (SNAP) [9, 7]. Ten, jako jeden z mála (další viz [7]) umožňuje paralelní běh. Je implementován v jazyce C, využívá POSIX vlákna a primitiva OpenMP. Je šířen jako open-source. Framework se mi podařilo zprovoznit pouze na systémech založených na OS Linux (konkrétně na Ubuntu 12.10 64bit a Debian 3.2.0).

V [7] je uvedeno, že oproti existujícím řešením je SNAP často 10-100 rychlejší a může pracovat se sítěmi majícími řádově až 10^9 vrcholů. Též zde naleznete algoritmy, v tomto frameworku použité (např. paralelní verze algoritmu BFS aj.). SNAP mj. umožňuje identifikaci komunit, výpočet C_B a centralit hran. Výpočet C_C bylo nutné doimplementovat. Taktéž jsem doimplementoval výpočet poloměru sítě a ASPL. Výpočty všech výše uvedených hodnot probíhají v jednom průchodu – na rozdíl od některých jiných programů, kde je v jednom průchodu možné vypočítat pouze jednu metriku. SNAP též umožňuje hodnoty centralit aproximovat, což považuji za velmi přínosné (viz dále). V [9] je uvedeno, že při 5% aproximaci se průměrná chyba u nejcentrálnějších vrcholů (nejvyššího 1%) pohybuje pod hranicí 20%. Veškeré výpočty jsou procesorově náročné a není-li SNAP omezen, využívá 100% procesorového času.

Pajek XXL a Gephi

Dalším programem, který jsem využil pro časově nenáročné výpočty (např. D_C, počet komponent apod.), byl Pajek XXL [32]. Ten též umožňuje práci s rozsáhlými sítěmi, nicméně neumožňuje paralelní běh. U programu Gephi (viz [11]) oceňuji velmi kvalitní vizualizaci sítě, se kterou je možné různě manipulovat (např. dle různých metrik zvětšovat a obarvovat vrcholy). Kvůli vizualizaci však není možné pracovat s rozsáhlými sítěmi (jako maximum se uvádí 50 tisíc vrcholů – tuto hodnotu se mi však podařilo překonat). Gephi též neumožňuje paralelní běh. Program je navíc velmi paměťově náročný a příležitostně se chová nestabilně. Lze ho však doporučit pro analýzu menších sítí.

6.2 Základní informace o síti

S použitím programů, popsaných v předešlé části, jsem získal pro *neorientovanou síť*[1] následující informace:

Tab. 6.1: Informace o neorientované síti.

Poč. vrcholů	Poč. hran	Vrch. v nej. komp.	Poč. komp.	Předpokl. % OSN
∼ 578 tisíc	∼ 6 788 tisíc	99,99 %	53	72-96%

Poloměr	ASPL	Hustota	Prům. d_G	Max. d_G	γ
15	4,87	4,06E-5	23,48	5254	3,5

Z tabulky 6.1 vyplývá, že mimo hlavní komponentu se v síti nachází dalších 52 komponent (každou tvoří jeden vrchol). To je pravděpodobně způsobeno existencí dupli-

[1] Vzhledem ke stahování pouze výstupních hran, velmi rychlou změnou OSN a možností vypnutí/zapnutí autorizace přátel, není snadné rozhodnout, který model je vhodnější. Proto jsou dále uvedeny oba modely.

citních profilů (viz část 4.4) a rychlou změnou dat v OSN. Vzhledem k její velikosti nemá tato skutečnost na výsledná data vliv.

Jelikož je hodnota $ASPL \leq 6,5$, podařilo se pro neorientovanou síť *potvrdit hypotézu šesti kroků* (průměrný počet prostředníků = $ASPL - 1$, neboť $ASPL = 1$ je přímý kontakt, viz část 2.3 a [29]). Zjištěná hodnota ASPL je velmi podobná (4,87 oproti 4,74) hodnotě, zveřejněné v práci [6], popisující strukturu OSN Facebook.com. Každého uživatele ve zkoumané síti je tak možné průměrně kontaktovat pomocí 4 prostředníků. Délka nejdelšího řetězce je pak 14 prostředníků (v experimentu prováděném Milgramem jich bylo 10, viz 2.3). Též se potvrdilo, že hodnota ASPL může být odhadnuta jako: $\ln(n)/\ln\ln(n)$. V našem případě tedy 5,13. ASPL v dalších OSN jsou taktéž podobné a to sice v intervalu 4,25 - 5,88, viz [30].

Tab. 6.2 uvádí informace získané z *orientované sítě*[2].

Tab. 6.2: Informace o orientované síti.

Poloměr	ASPL	Hustota	Prům. d_G^-, d_G^+	Max. d_G^+	Max. d_G^-
17	5,07	2,03E-5	11,74	5254	590

γ vstup.	γ výstup.	Vzájemných hran
3,5	3,5	31,5%

Oproti OSN Orkut, či jiným, je průměrný počet hran (prům. d_G) téměř 10x menší (viz [30]). Za zajímavý fakt považuji to, že pouze jedno ze tří přátelství je vzájemné. To je v porovnání s jinými OSN velmi málo. Např. Flickr má údajně tuto hodnotu 62%, LiveJournal 73,5% a YouTube 79,1% [30]. To může naznačovat, že se ve zkoumané síti vyskytuje poměrně velké množství slabých vazeb nebo to, že se ve skutečnosti někteří uživatelé navzájem neznají.

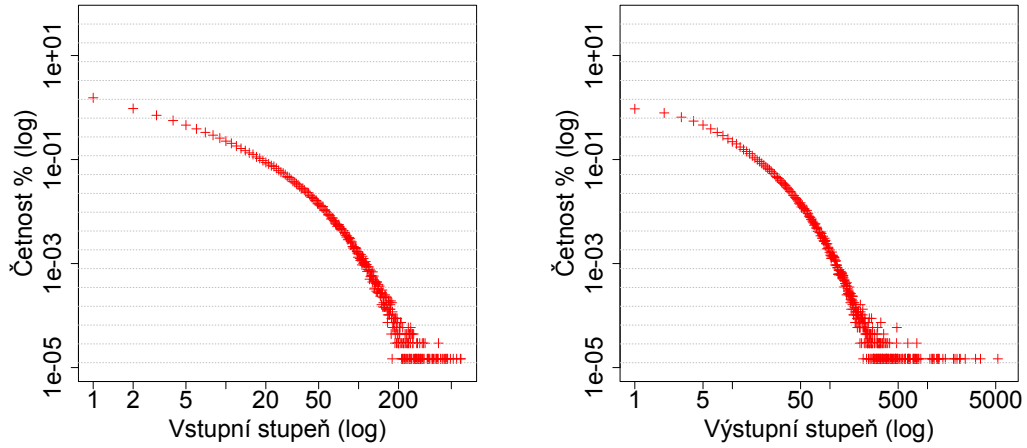
Maximální výstupní stupeň je téměř desetkrát větší než maximální vstupní stupeň. Z toho vyvozuji, že se v síti nenachází extrémně populární či prestižní osoby („pouze“ 50 krát častěji vyhledávané nežli je průměr). Naopak zde existuje několik osob shromažďujících velký počet kontaktů (jejich počet kontaktů přesahuje až 500 krát průměr).

Vysoký exponent konektivity[3] pravděpodobně umožňuje existenci kritického prahu [10]. Pokud bychom náhodně odebírali vrcholy, síť by se nejspíše rozpadla na více komponent. Ztrácí se tak jedna z výhod bezškálových sítí.

[2] ASPL a poloměr byly získány 20% aproximací, viz dále.

[3] Vypočten pomocí nástroje dostupného na `http://tuvalu.santafe.edu/~aaronc/powerlaws` metodou maximální věrohodnosti s hodnotou Kolmogorov–Smirnova testu rovné 0,0297. Tato malá hodnota naznačuje poměrně přesnou aproximaci [30].

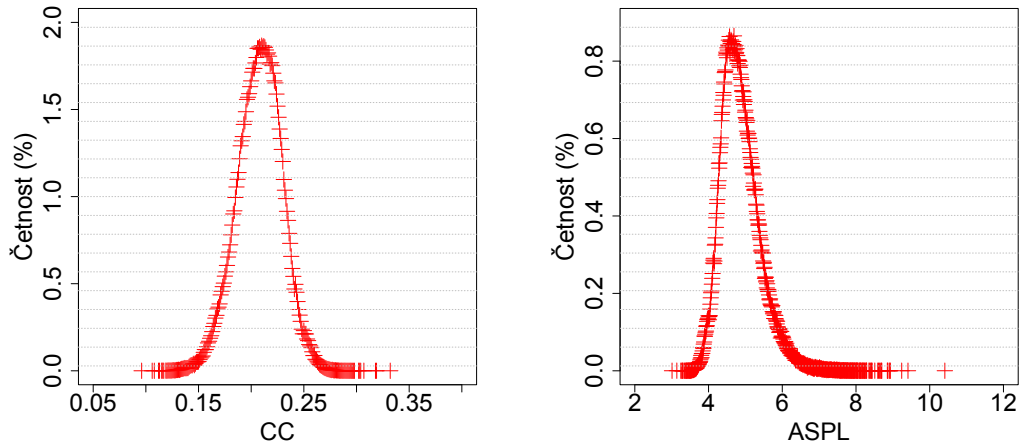
Na obr. 6.2 naleznete četnosti vstupního (obr. 6.2a) a výstupního stupně (obr. 6.2b).



(a) vstupní – d_G^- (b) výstupní – d_G^+

Obr. 6.2: Četnosti vstupního a výstupního stupně.

Na obr. 6.3 naleznete výskyt četností hodnot C_C (obr. 6.3b) a ASPL (obr. 6.3a). I když se rozdělení stupně vrcholů neřídí normálním rozdělením pravděpodobnosti tyto dvě téměř ano.



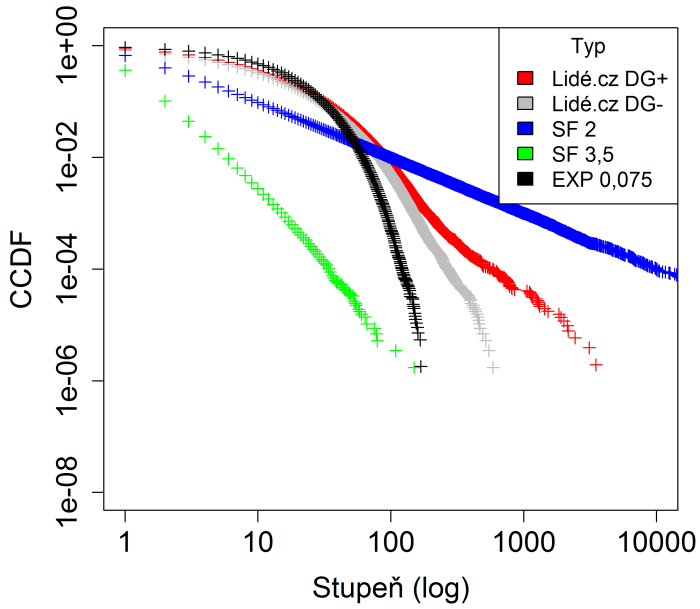
(a) C_C (b) ASPL

Obr. 6.3: Četnosti hodnot C_C a ASPL.

Z obr. 6.4[4] vyplývá, že i přes poměrně dobrou aproximaci koeficientu konektivity (γ), je v této síti tlumení počtu hran (zejména u vrcholů s menším stupněm) menší než

[4]CCDF – doplňková kumulativní distribuční funkce, vypočtena jako $1 - F(x)$, kde $F(x)$ je distribuční funkce.

je pro bezškálové sítě obvyklé. Jako EXP 0,075 je označeno CCDF exponenciálního rozdělení s parametrem $\lambda = 0.075$. Jako SF 2 a SF 3,5 jsou označeny CCDF standardních bezškálových sítí s příslušným koeficientem konektivity. Tvar CCDF OSN Lidé.cz (zejména d_G^+) připomíná CCDF vytvořenou ze sítě spolupráce herců získanou z IMDb[5] [5]. U té též nastupuje mocninný zákon až po určitém stupni vrcholů. Domnívám se tak, že v této síti je většina uživatelů schopna z počátku „sesbírat" několik kontaktů poměrně snadno[6]. Tato jejich schopnost se s přibývajícím stupněm snižuje a nastupuje mocninný zákon, který umožní dosáhnout velkého stupně pouze několika vrcholům.



Obr. 6.4: Porovnání několika různých CCDF.

6.3 Výsledky Centrality measures a jejich porovnání

Kvůli výpočetní složitosti (viz část 2.4) a časové náročnosti byly výpočty centralit prováděny na serveru `students.kiv.zcu.cz`. Na tom se nachází: 8x Intel(R) Xeon(TM) 3.0 GHz (dvou jádrové, tedy celkem 16x CPU) a 32GB RAM. Z časových důvodů (viz dále) jsem pro neorientovanou síť provedl 5% aproximaci a přesný výpočet. Pro orientovanou síť pak 5% a 20% aproximaci. Výpočet 5% aproximace trval cca 5,5 hodiny, 20% aproximace přibližně jeden den a přesný výpočet cca 4,5 dne.

[5] `http://www.imdb.com`

[6] Což vzhledem k tomu, že člověk má údajně 200-5000 sociálních vazeb [10, 29] není překvapivé.

Výsledky hodnot orientované i neorientované sítě, jsou si v našem případě velmi podobné (viz tab. v příloze E a tab. 6.3). Aproximace též dosahují velmi podobných výsledků.

Vzhledem k malému růstu ASPL v neorientované síti (cca 2% při 15% rozšíření vrcholů, viz tab. D.1) lze předpokládat, že *je hypotéza šesti kroků splněna i pro orientovanou síť* a to i přes to, že je pouze 31,5% přátelstvích vzájemných. Předpokládám tedy, že je síť dobře propojena, zejména mezi významnými vrcholy, které tvoří tzv. jádro sítě (viz příloha F).

V tab. 6.3 naleznete žebříčky uživatelů, kteří se dle základních metrik CM (tj. C_C, C_B a C_D) umístili na prvních dvaceti místech. Žebříčky byly sestaveny pro přesné hodnoty CM pro neorientovanou síť. Uživatelé, vyskytující se v první desítce, mají napříč žebříčky přiřazenou jednotnou barvu pro snadnější identifikaci. Všimněte si, že i přes různé způsoby výpočtů těchto metod se v první desítce téměř všechna jména opakují a jsou na podobných pozicích.

Tab. 6.3: Uživatelé OSN seřazeni dle dosaženého místa pro jednotlivé metriky CM – přesný výpočet neorientované sítě.

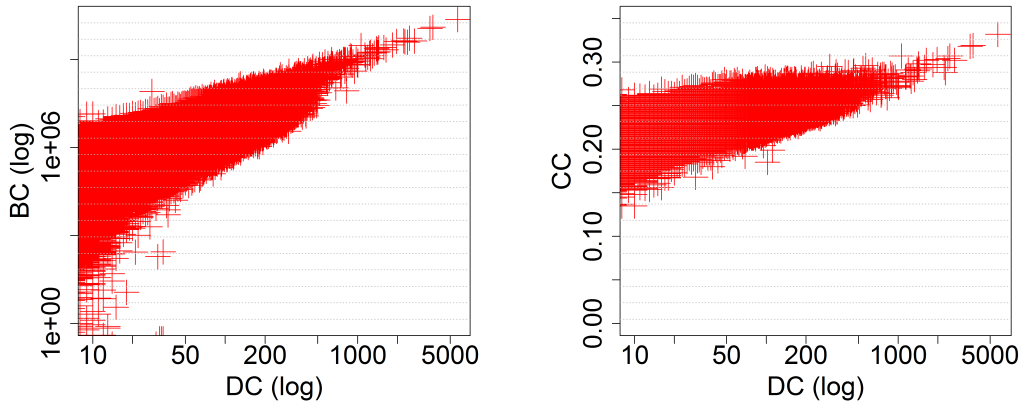
#	C_B	C_C	C_D
1	honzek098	honzek098	honzek98
2	sochurek.j	sochurek.j	sochurek.j
3	r.hasek	r.hasek	r.hasek
4	melly.19	nezkrotny.dablicek.lucie.xd	martinoof
5	ashraf.shafeek	melly.19	melly.19
6	martinoof	flamez10	flamez10
7	lukas.klement@post.cz	lukas.klement@post.cz	vvladimirvanek
8	vvladimirvanek	martinoof	ashraf.shafeek
9	flamez10	vivi.elien.186	lukas.klement@post.cz
10	durdil.d	kapradorosty	durdil.d
11	nezkrotny.dablicek.lucie.xd	durdil.d	vivi.elien.186
12	topol.d	marcela-marsi	topol.d
13	zasova1991	nikollka22	zasova1991
14	rocker.k	topol.d	kapradorosty
15	ajik-1	ashraf.shafeek	ajik-1
16	vivi.elien.186	ajik-1	martin.karatista
17	kapradorosty	s.e.x.o.n.t.h.e.b.e.a.c.h	rocker.k
18	martin.karatista	rocker.k	citronek.zx
19	kubasvitak	zasova1991	avoldies
20	citronek.zx	beruska.no1	petrlupik

Na obr. E.1 naleznete tytéž žebříčky pro 5% aproximaci sítě (aproximace vychází z toho, že nás zajímají pouze vrcholy s nejvyšší hodnotou – tj. centra). Již po 5% aproximaci, která je 20 krát rychlejší než přesný výpočet, metody C_B i C_C (D_C není výpočetně náročná, nemá ji tedy smysl aproximovat) poměrně přesně centra odhalí.

Obr. D.2 zobrazuje grafické porovnání aproximovaných hodnot s hodnotami přesnými. Z obr. D.2b a D.1b, zobrazující pouze nejvyšší hodnoty, vyplývá, že většina center je umístěna na správném místě (body tvoří téměř přímku). Aproximace ve frameworku SNAP byla optimalizována pro metodu C_B (viz. [9]), proto je její aproximace přesnější. Aproximace C_C ve frameworku SNAP původně implementována nebyla, přesto vysoké hodnoty aproximuje poměrně dobře.

V tab. E.2 a E.3 naleznete tytéž žebříčky pro síť orientovanou. Aby bylo možné výpočty všech metrik provést v jednom běhu, byla namísto metriky C_C použita metrika C_C^- (blízkost dle vstupních hran) – tzn., jak snadno se k vrcholu dostanou vrcholy ostatní [32]. Z toho důvodu ve výše uvedených tabulkách naleznete metriku C_D^- a žebříček se od dříve uvedených poněkud liší. C_B je opět aproximována dobře. C_C^- již příliš ne z důvodu, který byl uveden dříve.

Obr. 6.5 zobrazuje závislosti mezi hodnotami CM ve zkoumané síti. Z obr. 6.5a a 6.5b vyplývá, že v této síti jsou hodnoty C_B a C_C úměrné hodnotě C_D. Tzn. čím větší je stupeň vrcholu, tím má větší hodnoty C_B a C_C (patrné zejména u vysokých hodnot).



(a) C_B x C_D (b) C_C x C_D

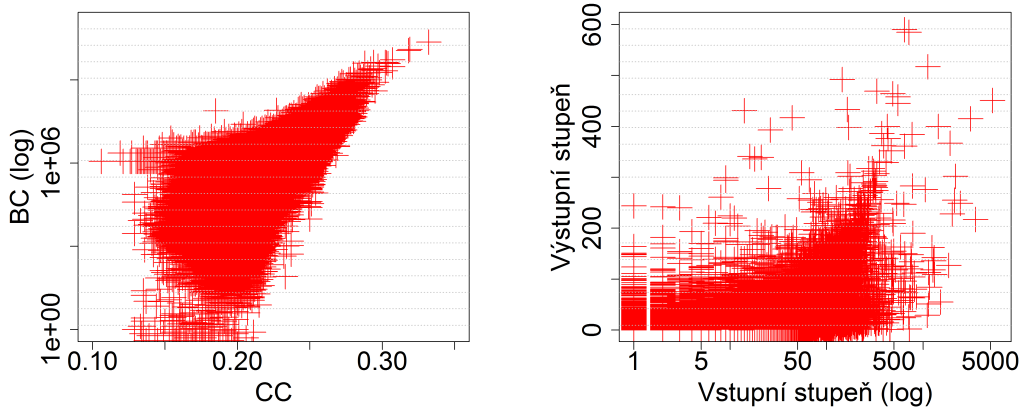
Obr. 6.5: Vzájemné závislosti centralit ve zkoumané síti pro přesný výpočet neorientované sítě.

Obecně tomu tak být nemusí. Pokud by např. síť byla tvořena více rozsáhlými komunitami, které by spojovalo pouze několik vrcholů, hodnota C_B těchto vrcholů by byla vysoká, nehledě na jejich stupeň.

Jelikož se zde takové vrcholy nevyskytují, lze vyloučit, že je síť rozdělena do velkých komunit (např. dle krajů, generací apod.), které jsou spojeny pouze několika vrcholy. Síť je naopak dobře propojená a při komunikaci, využívající nejkratších cest, jí většina informací putuje přes jádro sítě (viz příloha F – podsíť vytvořená z uživatelů mající více než 500 hran).

Obr. 6.6a ukazuje, že v síti existuje několik vrcholů majících malou hodnotu C_C, ale relativně velkou hodnotu C_B. Z toho usuzuji, že v síti existují vrcholy (či malé komunity), které jsou na „okraji“ sítě a k jejímu zbytku jsou připojeny např. pouze prostřednictvím jednoho či dvou vrcholů, které jsou od ostatních vrcholů „vzdálené“. Lze očekávat, že k těmto vrcholům se dostanou informace ze sítě mezi posledními případně vůbec.

Obr. 6.6b znázorňuje nezávislost vstupního a výstupního stupně. V této síti tak existují vrcholy, které jsou populární (vyhledávané s vysokým vstupním stupněm), ale nejsou příliš vlivné (nemají vysoký výstupní stupeň) a naopak.



(a) C_B x C_C (b) C_D^- x C_D^+

Obr. 6.6: Vzájemné závislosti centralit ve zkoumané síti pro přesný výpočet neorientované sítě.

Vzhledem k tomu, že je tato síť dobře propojená, lze pro určování významnosti vrcholů doporučit metriku C_D, která není příliš výpočetně náročná. Případně používat 5% aproximaci metrik C_B a C_C, které jsou pro odhalení nejvýznamnějších vrcholů (zejména u neorientované sítě) dostatečně přesné.

7 Závěr

Práce si kladla za úkol seznámit čtenáře s některými metodami, které analýza sociálních sítí nabízí. Za tímto účelem byla analyzována rozsáhlá sociální síť přátel, kterou se podařilo získat z online sociální sítě (Online social network – OSN) Lidé.cz. Ukázalo se, že provozovatel této OSN se automatickému procházení nebrání a podařilo se tak projít více než milion webových stránek. Z těchto stránek byla vytvořena databáze, obsahující přátelství mezi uživateli a informace o uživatelích. Odhaduji, že tato databáze v době, kdy byla aktuální, obsahovala 72-96% uživatelů, kteří se v této OSN nacházeli.

V teoretické části byl představen model bezškálové sítě, kterým se dle Barabásiho (viz [10]) řídí většina reálných komplexních sítí. Získanou síť (ta byla tvořena téměř 600 tisíci vrcholy a 7 miliony hran) jsem se oproti tomuto modelu pokusil verifikovat. Ukázalo se, že se dle tohoto modelu „řídí“ pouze částečně. Dále bylo zjištěno, že se její topologie podobá sociální síti spolupráce mezi herci.

Byly představeny základní metody, které jsou schopné odhalit uživatele, mající ve zkoumané síti klíčovou roli. Podle těchto metod byly sestaveny žebříčky klíčových uživatelů v síti, které jsou mezi sebou porovnány. Toto porovnání odhalilo, že v síti nejsou lidé rozděleni do velkých komunit (např. dle generací, pohlaví, bydliště apod.), které by spojovalo pouze několik uživatelů. Naopak je síť velmi dobře propojená a nejvíce využívanými cestami jsou ty mezi centry.

Jelikož jsou si získané žebříčky velmi podobné, lze pro identifikaci center v této síti doporučit nejrychlejší, i když obecně nejméně přesnou, metodu – Freemanovu degree centrality. Její výhodou je okamžitý výpočet. Přesný výpočet dalších metrik, vzhledem k jejich výpočetní náročnosti a velikosti sítě, zabral 4,5 dne. Obecně tak lze doporučit 5% aproximaci dalších metrik (Closeness a Betweenness centrality). Jak bylo ukázáno, výsledky těchto aproximací jsou dostatečně přesné.

Díky dobře propojenému jádru má síť velmi malou průměrnou délku nejkratší cesty a to sice 4,87. Pro orientovaný model je tato hodnota o několik procent vyšší. Stále je však menší, než hodnota uváděná Milgramem (tj. 6,5). Pro tuto sociální síť přátel se tak podařilo ověřit hypotézu šesti kroků. Získaná hodnota je navíc velmi podobná hodnotám, které jsou uváděné v podobných výzkumech, zkoumajících jiné OSN.

Práce se též zabývala analýzu dat v OSN uložených. Z těch byly vytvořeny statistické výstupy, které byly porovnány s podobnými průzkumy, aby bylo možné posoudit jejich relevantnost. Z těchto informací je možné např. zjistit, jak velkou potencionální skupinu může reklamní sdělení oslovit a mohou tak být použity reklamními či marketingovými agenturami. Též poskytují jasnější informace o sociodemografii této OSN (oficiální údaje, uváděné na OSN Lidé.cz, jsou tvořeny ze vzorku tvořícího cca 0,5% mnou analyzovaných dat). V OSN Lidé.cz jsou konkrétně nejvíce zastoupenou

skupinou lidé ve věku 15-29 let, se zájmy o sport (zejména cyklistiku a kolektivní míčové hry) a o hudbu (zejména rock a pop). Nejpopulárnější je tato OSN u uživatelů v Plzeňském kraji.

Dalšími zkoumanými vlastnostmi byla vzdělanost, znalost cizích jazyků a problematika konzumace alkoholu a kouření u mladistvých. Získané údaje se, s poměrně malou tolerancí, shodují s údaji uváděnými v jiných zdrojích. Mohou tak být považovány za relevantní a analýzu obsahu OSN lze doporučit jako alternativní formu k formě „dotazníkové“ a ušetřit tak náklady, které jsou s ní spojené.

7.1 Budoucí práce

Jako možné rozšíření se nabízí analýza komunit, která zkoumá a identifikuje komunity, ve kterých se většina uživatelů navzájem zná. Též by bylo možné analyzovat dynamiku sítě. S drobnými úpravami databáze a webových robotů by bylo možné uchovávat různé verze sítě a zkoumat, jak se časem měnila. Výsledky této analýzy by mohli např. odhalit vzory, jak se virtuální přátelství navazují, mění či ruší.

Dále by bylo možné získat více informací vycházejících z dat v OSN obsažených. Např. je možné zkoumat různé závislosti mezi bydlištěm a zálibami uživatelů. Tyto otázky, zaměřené na lokalitu, by mohli být přínosné např. pro lokální obchodníky, kina či organizátory sportovních akcí.

Zajímavou možností by též bylo porovnání výsledků získaných pomocí základních (tj. Freemanových) metod Centrality measures s metodami dalšími (např. s metodami PageRank, Bonachii degree či HITS).

V této práci jsem pro zjednodušení, urychlení výpočtu a urychlení získání dat, považoval všechny hrany za stejně významné. Ve skutečnosti tomu tak být nemusí. Hranám lze přiřadit různé váhy. Ty mohou být získány např. pomocí kombinace geografické vzdálenosti uživatelů, jejich společnými zájmy, počtu a obsahu zpráv, které si píší na „nástěnku“, jaký komentář uvádějí u popisu přátelství a je-li přátelství vzájemné. Poté může být zajímavé pozorovat, zda a jak se první místa v žebříčcích změní.

Posledním návrhem na budoucí práci je možný průzkum, z jakého důvodu jsou centra, či konkrétní uživatelé, centry. Mají nějaké společné zájmy či vlastnosti? Znají stejné uživatele? Jak by asi vypadal a jaké by měl mít vlastnosti „ideální“ uživatel, proto aby se mohl stát populárním centrem a z této pozice profitovat?

Literatura

[1] Zákon č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon).

[2] *The Web Robots Pages* [online]. [cit. 28.4.2013]. Dostupné z: `http://www.robotstxt.org`.

[3] *Scrapy* [online]. [cit. 28.4.2013]. Dostupné z: `http://scrapy.org`.

[4] ČADA, R. – RYJÁČEK, Z. – KAISER, T. *Diskrétní matematika*. Západočeská univerzita, 2004.

[5] ALBERT, R. – BARABÁSI, A.-L. Statistical mechanics of complex networks. *Reviews of modern physics*. 2002, 74, 1, s. 47.

[6] BACKSTROM, L. et al. Four degrees of separation. *arXiv preprint arXiv:1111.4570*. 2011.

[7] BADER, D. A. *Parallel Programming for Graph Analysis* [online]. [cit. 28.4.2013]. Dostupné z: `http://www.cc.gatech.edu/~bader/papers/PPoPP12`.

[8] BADER, D. A. – MADDURI, K. Parallel algorithms for evaluating centrality indices in real-world networks. In *Parallel Processing, 2006. ICPP 2006. International Conference on*, s. 539–550. IEEE, 2006.

[9] BADER, D. A. – MADDURI, K. SNAP, Small-world Network Analysis and Partitioning: an open-source parallel graph framework for the exploration of large-scale networks. In *Parallel and Distributed Processing, 2008. IPDPS 2008. IEEE International Symposium on*, s. 1–12. IEEE, 2008.

[10] BARABÁSI, A.-L. *V pavučině sítí*. Paseka, 2005. ISBN 80-7185-751-3.

[11] BASTIAN, M. – HEYMANN, S. – JACOMY, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. *International AAAI Conference on Weblogs and Social Media*. 2009. Dostupné z: `http://gephi.org`.

[12] BATAGELJ, V. *Friendship and unionization in a hi-tech firm.* [online]. [cit. 28.4.2013]. Dostupné z: `http://vlado.fmf.uni-lj.si/pub/networks/data/esna/hiTech.htm`.

[13] BRANDES, U. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*. 2001, 25, 2, s. 163–177.

[14] BUŠTÍKOVÁ, L. Analýza sociálních sítí. *Sociologický časopis*. 1999, 35, 2, s. 193–206.

[15] CATANESE, S. et al. Extraction and analysis of facebook friendship relations. *Computational Social Networks*. 2012, s. 291–324.

[16] DE NOOY, W. – MRVAR, A. – BATAGELJ, V. *Exploratory social network analysis with Pajek*. Cambridge University Press, 2011. ISBN 9780521602624.

[17] ELLISON, N. B. et al. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*. 2007, 13, 1, s. 210–230.

[18] FERRARA, E. – FIUMARA, G. *Mining and Analysis of Online Social Networks*. University of Messina, 2012.

[19] FRANĚK, M. – MUŽÍK, P. *Hudební preference a její souvislost s některými osobnostními rysy* [online]. [cit. 28.4.2013]. Dostupné z: `http://acta.musicologica.cz/06-03/0603s02.html`.

[20] FREEMAN, L. C. A set of measures of centrality based on betweenness. *Sociometry*. 1977, s. 35–41.

[21] GREGOROVIČ, T. *Extrakce informací ze sociálních médií*. Masarykova Univerzita, 2010.

[22] HANKE, T. *Import dat ze služby Google Scholar do formátu XML*. Západočeská univerzita, 2012.

[23] HANNEMAN, R. A. – RIDDLE, M. *Introduction to social network methods*. University of California Riverside, 2005.

[24] JARUŠEK, P. *Analýza sociální sítě organizátorů zážitkových akcí*. Masarykova Univerzita, 2008.

[25] JOHNSON, J. A. et al. *Social Network Analysis: A Systematic Approach for Investigating* [online]. [cit. 28.4.2013]. Dostupné z: `http://www.fbi.gov/stats-services/publications/law-enforcement-bulletin/2013/March/social-network-analysis`.

[26] ŠLERKA, J. *Social network analysis pro začátečníky* [online]. [cit. 28.4.2013]. Dostupné z: `http://www.lupa.cz/clanky/social-network-analysis-pro-zacatecniky`.

[27] LESKOVEC, J. Social media analytics: tracking, modeling and predicting the flow of information through networks. In *Proceedings of the 20th international conference companion on World wide web*, s. 277–278. ACM, 2011.

[28] LONG, J. *Google Hacking*. Zoner Press, 2005. ISBN 80-86815-31-5.

[29] MILGRAM, S. The small world problem. *Psychology today*. 1967, 2, 1, s. 60–67.

[30] MISLOVE, A. et al. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, s. 29–42. ACM, 2007.

[31] MRAVČÍK, V. Evropská školní studie o alkoholu a jiných drogách (ESPAD). In *Zaostřeno na drogy 1/2012*. Úřad vlády ČR, 2012.

[32] MRVAR, A. *Network Analysis using Pajek* [online]. [cit. 28.4.2013]. Dostupné z: `http://mrvar.fdv.uni-lj.si/sola/info4`.

[33] NOSÁL, P. *Vzdělávání dospělých* [online]. [cit. 28.4.2013]. Dostupné z: `http://www.czso.cz/csu/tz.nsf/i/prezentace_z_tiskove_konference_vzdelavani_dospelych/$File/csu_tk_vzdelavani_prezentace.pdf`.

[34] SAK, P. *Stav a vývoj znalostí cizích jazyků české populace* [online]. [cit. 28.4.2013]. Dostupné z: `http://www.insoma.cz/index.php?id=1&n=1&d_1=paper&d_2=jazyky_cz`.

[35] SEZNAM.CZ. *Lidé.cz – Nevhodné chování a zakázané aktivity* [online]. [cit. 28.4.2013]. Dostupné z: `http://napoveda.seznam.cz/cz/nevhodne-chovani-a-zakazane-aktivity.html`.

[36] SEZNAM.CZ. *Smluvní ujednání - Seznam účet* [online]. [cit. 28.4.2013]. Dostupné z: `http://registrace.seznam.cz/licenceScreen`.

[37] STERLY, R. *Výsledky ankety Sport roku 2012 – napínavý souboj mezi cyklistikou a fotbalem* [online]. [cit. 28.4.2013]. Dostupné z: `http://info.sportcentral.cz/blog/vysledky-ankety-sport-roku-2012-souboj-mezi-cyklistikou-a-fotbalem`.

Slovník pojmů a zkratek

ASPL	Průměrná délka nejkratší cesty (Average shortest path length)
BFS	Prohledávání do šířky (Breadth-first search)
CCDF	Doplňková kumulativní distribuční funkce (Complementary cumulative distribution function)
CM	Centrality measures
CSV	Comma-separated values
ČSÚ	Český statistický úřad
DFS	Prohledávání do hloubky (Depth-first search)
DOM	Document Object Model
GPL	General Public License
HTML	HyperText Markup Language
IBMd	Internet Movie Database
OS	Operační systém
OSN	Online sociální síť (Online social network)
RU_M	Měsíčních unikátních přístupů
SNAP	Small-world Network Analysis and Partitioning
URL	Uniform resource locator
XML	Extensible Markup Language
XPath	XML Path Language

Znaky použité ve vzorcích

Δ	Hustota grafu
γ	Exponent konektivity
C_B	Betweenness centrality
C_C	Closeness centrality
C_C^-	Closeness centrality dle vstupních hran
C_D	Degree centrality
C_D^+	Degree centrality dle výstupních hran
C_D^-	Degree centrality dle vstupních hran
D	Poloměr grafu
d_G	Stupeň vrcholu

Obsah DVD

Na přiloženém DVD naleznete následující složky a soubory:

- Data
  - lide-sit.net – síť ve formátu .net (viz příloha B)
  - lide-sit.gr – síť ve formátu .gr (viz příloha B)
  - lide-sit.sql – SQL Dump databáze se získanými daty
  - lide-struktura.sql – SQL Dump struktury databáze (viz část 4.2.1)
  - vypocty – složka obsahující vypočtené hodnoty CM
- Roboti
  - SNABot-sit – robot pro průchod a zpracování sítě přátelství
  - SNABot-profily – robot pro průchod a zpracování profilů
  - SNABot-centra – robot pro průchod a zpracování center
  - Scrapy.pdf a Scrapy-0.14.4.tar.gz – manuál a knihovna Scrapy
- Analyza
  - SNAP – rozšířený framework SNAP
  - statistiky – tabulky a prezentace uváděné ČSÚ
- R
  - csv – CSV soubory sloužící k tvorbě grafů
  - skripty – R skripty
  - pdf – PDF soubory s grafy vygenerované pomocí R skriptů
- SNABot-nastroje.jar – program umožňující generování CSV souborů pro tvorbu statistik a generování souboru se sítí
- BP-Marek-Naggy.pdf – elektronická verze BP
- BP-Marek-Naggy.rar – zdrojové dokumenty elektronické verze BP

Přílohy

A Uživatelské příručky

A.1 Webový robot SNABot

Pro spuštění webového robota je třeba mít nainstalovaný Python verzi 2.6 nebo 2.7 a knihovnu Scrapy. Tu můžete nalézt v [3] (průběh instalace je popsán v oficiální dokumentaci v části *Installation guide*). Dále je nutné vytvořit databázi s předepsanou strukturou (viz část 4.2.1 či SQL Dump na přiloženém DVD). Roboti byli vyvíjeni a testováni na OS Windows 7. Vzhledem k tomu, že všechny použité části jsou platformě nezávislé, mělo by být možné roboty spustit i na jiných OS. Před spuštěním je třeba v souboru `scrapy.bat` nastavit cestu, ve které je Python nainstalován.

Pro zahájení stahování spusťte příkazový řádek a přepněte se do složky, ve které se robot nachází. Toho následně spustíte příkazem:

```
scrapy crawl jmeno -s JOBDIR=adresar
```

Kde `jmeno` je jméno webového robota. Těmi jsou: SNABot-profily – pro průchod profilů, jejichž jména je nutné získat pomocí robota SNABot-sit –, SNABot-centra – pro průchod center, jejichž jména je nutné získat pomocí robota SNABot-sit – a SNABot-sit – pro průchod stránek moji-pratele, obsahující přátele uživatele.

Parametr `adresar` je adresář, do kterého budou ukládána data, díky kterým bude možné robota pozastavit a při dalším spuštění na toto stahování navázat. Nedoporučuji spouštět více robotů najednou. Jinak může dojít k jejich odpojení či nesprávné funkčnosti.

Pro přerušení stahování stiskněte v příkazové řádce kombinaci kláves `ctrl` + `c`. Robot bude po krátké době ukončen. Případně lze ukončení vynutit dvojitým stisknutím této kombinace. Poté však není zaručeno, že budou zpracovány všechny stránky korektně.

Pro navázání na přerušené stahování opět zadejte výše uvedený příkaz.

Roboti logují své akce do souborů `log-rrrr-mm-dd.txt`. V těch je možné odhalit případné problémy. Též jsou v nich uvedeny základní statistiky o stahování, jako je jeho rychlost, velikost přenesených dat apod.

A.2 Framework SNAP

Framework SNAP se mi podařilo spustit pouze na systémech založených na OS Linux. Před samotným spuštěním je třeba framework nakonfigurovat. Toho docílíte příkazem `./configure --enable-openmp`. Bez zadání parametru `-openmp` nebude umožněn paralelní výpočet! Po konfiguraci zadejte příkaz `make`, který zdrojové soubory přeloží (překlad může trvat i několik minut).

Pro spuštění výpočtu centralit přejděte do složky `test`. V terminálu poté zadejte následující příkaz:

```
./eval_vertex_betweenness -infile vstup.gr -outfile vystup.txt
   [-approx X]
```

Kde `vstup.gr` je soubor se sítí, která má být analyzována, ve formátu `.gr` (viz příloha B). `vystup.txt` je soubor, do kterého budou výsledky uloženy. Pomocí volitelného parametru `-approx` je možné výsledky aproximovat (viz část 6.3) a výpočet tak značně urychlit. `X` značí kolika procentní aproximací bude výpočet proveden (nejčastěji se uvádí 5%).

Přesný výpočet může být časově velmi náročný. Při cca 580 tisících vrcholech trval na 16 jádrovém procesoru cca 4,5 dne. Pokud není program omezen, využívá po čas výpočtu 100% procesorového času.

A.3 SNABot-nastroje

Tento nástroj slouží k odstranění duplicit, které vznikly z důvodu uvedeném v části 4.4. Dále umožňuje vygenerování souboru, obsahujícího sociální síť a generování CSV souborů, které umožňují automatickou aktualizaci grafů uvedených v kapitole 5. Pro správnou funkčnost je třeba, aby data byla uložena v databázi mající dříve uvedenou strukturu (viz část 4.2.1 či SQL Dump na přiloženém DVD). Soubory obsahující síť je možné vygenerovat ve dvou formátech (viz příloha B).

Pro odstranění duplicit a vygenerování souboru se sítí zadejte do příkazové řádky následující příkaz:

```
java -jar SNABot-nastroje.jar -d/n -format [-csv]
```

Kde `-d` je příznak pro odstranění duplicit (těch se v databázi nachází cca 10 tisíc). Při zadání tohoto příznaku může zpracování trvat několik minut. Z důvodu zpětné

kompatibility nebyly duplicity odstraněny přímo v databázi. Při zadání příznaku `n` duplicity odstraněny nebudou a soubory jsou vygenerovány téměř ihned.

Parametr `-format` určuje, v jakém formátu bude síť uložena. Na výběr jsou tři možnosti: `-net` – síť bude možné analyzovat pomocí většiny programů (jako jsou např. Pajek XXL, Gephi atd.), `-gr` – síť bude možné analyzovat pomocí frameworku SNAP, který umožňuje paralelní výpočet metrik C_C a C_B, `-oba` – budou vytvořeny oba soubory s výše uvedenými formáty.

Pokud bude zadán volitelný parametr `-csv`, z databáze budou vygenerovány CSV soubory, které umožňují automatickou tvorbu grafů pomocí statistického nástroje R (R skripty naleznete na přiloženém DVD ve složce `R`).

B Formáty souborů

Síť lze uložit do dvou různých formátů, které budou dále popsány.

Formát .gr

Soubor s příponou `.gr` je používán frameworkem SNAP při výpočtu centralit C_C a C_B. Reprezentace sítě má následující formát:

```
p 578119 6788208 u u 1
1 1211
......
578119 22513
```

První řádka definuje model, se kterým se bude dále pracovat, ostatní řádky tvoří jednotlivé hrany. `p` definuje první řádku, číslo za ním určuje počet vrcholů. Druhé číslo určuje počet hran. Následující písmeno značí: `u` – síť je neorientovaná, `d` – síť je orientovaná. Třetí písmeno určuje zda je síť vážená – `w`, či nevážená – `u`. Váhy je možné přidat na konec řádek tvořící hrany např. takto: `1 1211 5`. Poslední číslo značí, zda je první vrchol číslován od jedničky, či od nuly. Dle toho je toto číslo nutné nastavit. U řádků, určujících hrany, je nutné dbát na to, aby se v nich nevyskytovali nepatřičné mezery (např. na konci řádek).

Formát .net

Tento formát je používán programy Gephi, Pajek XXL a některými dalšími programy. Soubor má následující formát:

```
*Vertices 578119
1 "jmeno"
.........
578119 "jmeno"
*Edgeslist
1 1211 1212 1213 1214
.........
578118 278044
```

Číslo za značkou `*Vertices` udává počet vrcholů. Vrcholy jsou číslovány od 1 a je možné je různě označit. V našem případě jménem. Značka `*Edgelist` říká, že se jedná o neorientované hrany. V případě hran orientovaných použijeme značku `*Arcslist`. Též je možné použít značky `*Edges` a `*Arcs`. V tom případě je však nutné, aby se na jedné řádce nacházela pouze jedna hrana. Hranám je též možné přiřadit váhu přidáním její hodnoty na konec řádky (např. `1 1211 5`).

C Rozšíření statistik

V této příloze naleznete tabulku s detailním rozdělením mladistvých do skupin podle toho, jak často údajně konzumují alkohol a kouří. Tabulka se vztahuje k části 5.4.

Tab. C.1: Skupiny kuřáků a konzumentů alkoholu u uživatelů mladších 18 let.

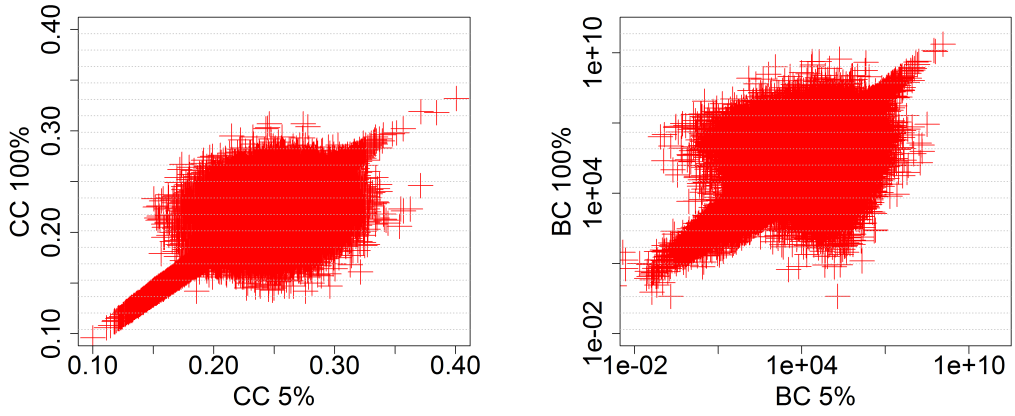
Konzumace alkoholu	Kouření	Mladistvých
Neuvádí	Neuvádí	6755
Neuvádí	Hodně	58
Neuvádí	Ne	3859
Neuvádí	Občas	216
Neuvádí	Pořád	42
Abstinent	Neuvádí	95
Abstinent	Hodně	66
Abstinent	Ne	4539
Abstinent	Občas	243
Abstinent	Pořád	67
Denně	Neuvádí	7
Denně	Hodně	99
Denně	Ne	51
Denně	Občas	31
Denně	Pořád	13
I bez míry	Neuvádí	115
I bez míry	Hodně	448
I bez míry	Ne	794
I bez míry	Občas	572
I bez míry	Pořád	91
Pořád	Neuvádí	3
Pořád	Hodně	448
Pořád	Ne	794
Pořád	Občas	572
Pořád	Pořád	91
S mírou	Neuvádí	424
S mírou	Hodně	610
S mírou	Ne	5544
S mírou	Občas	1860
S mírou	Pořád	282
	Celkem	27048

D Porovnání výsledků aproximace s přesným výpočtem

V této příloze, která se váže k části 6.2, naleznete porovnání hodnot CM, poloměru sítě a ASPL. Též jsou zde porovnány přesné výsledky a výsledky získané pomocí aproximace. V tabulce D.1 naleznete hodnoty poloměru sítě a ASPL pro různé typy výpočtů. Na obrázcích naleznete porovnání nakolik jsou aproximované hodnoty (resp. umístění jednotlivých vrcholů) shodné s přesným výsledkem. V ideálním případě by se mělo jednat o přímku. Je zde uvedeno porovnání 5% aproximace a přesného výpočtu neorientované sítě. Obr. D.2 zobrazuje celkové porovnání. Na následující stránce naleznete detail tohoto porovnání pro vysoké hodnoty. Čili těch, které nás většinou zajímají.

Tab. D.1: Porovnaní ASPL a poloměru sítě pro různé typy výpočtů.

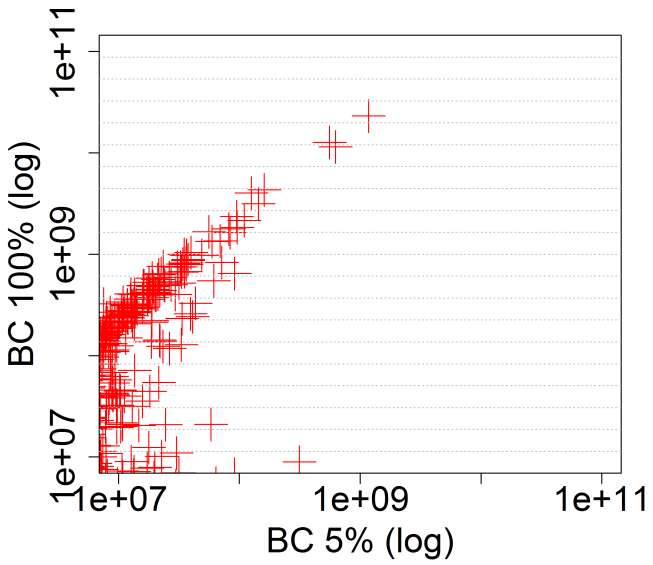
Typ sítě	Aproximace	ASPL	Poloměr
Neorientovaná	5%	4,39	12
Neorientovaná	100%	4,86	15
Orientovaná	5%	4,97	15
Orientovaná	20%	5,07	17



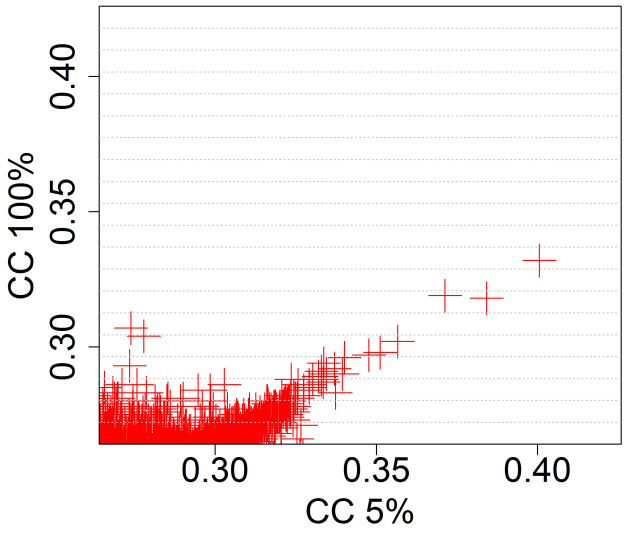
(a) C_C – všechny hodnoty

(b) C_B – všechny hodnoty

Obr. D.1: Porovnání 5% aproximace a přesného výpočtu pro neorientovanou síť.



(a) C_B – detail vysokých hodnot



(b) C_C – detail vysokých hodnot

Obr. D.2: Porovnání 5% aproximace a přesného výpočtu pro neorientovanou síť.

E Výsledky Centrality measures pro orientovanou síť a aproximace

V této příloze naleznete žebříčky uživatelů, kteří se umístily na prvních 20 místech, dle základních metrik Centrality measures[1]. Uživatelé, vyskytující se na prvních 10 místech v žebříčku uvedeném v části 6.3, mají pro snadnější identifikaci napříč žebříčky přiřazenou jednotnou barvu.

Jsou zde uvedeny žebříčky pro síť orientovanou i neorientovanou a jejich 5% a 20% aproximace. Přesný výpočet pro neorientovanou síť naleznete v tabulce 6.3.

Tab. E.1: Uživatelé OSN seřazeni dle dosaženého místa pro jednotlivé metriky CM – 5% aproximace neorientované sítě.

#	C_B	C_C	C_D
1	honzek098	honzek098	honzek098
2	r.hasek	r.hasek	sochurek.j
3	sochurek.j	sochurek.j	r.hasek
4	melly.19	melly.19	martinoof
5	flamez10	flamez10	melly.19
6	martinoof	martinoof	flamez10
7	lukas.klement@post.cz	vivi.elien.186	vvladimirvanek
8	nezkrotny.dablicek.lucie.xd	nezkrotny.dablicek.lucie.xd	ashraf.shafeek
9	ashraf.shafeek	lukas.klement@post.cz	lukas.klement@post.cz
10	durdil.d	kapradorosty	durdil.d
11	vvladimirvanek	durdil.d	vivi.elien.186
12	vivi.elien.186	janmukarovsky	topol.d
13	topol.d	ajik-1	zasova1991
14	kapradorosty	topol.d	kapradorosty
15	rocker.k	nikollka22	ajik-1
16	ajik-1	marcela-marsi	martin.karatista
17	zasova1991	martin.karatista	rocker.k
18	zlatahela50@email.cz	lusie44	citronek.zx
19	kolousek2@email.cz	karollecka	avoldies
20	martin.karatista	s.e.x.o.n.t.h.e.b.e.a.c.h	petrlupik

[1]Legenda k tabulkám: C_B – Betweenness centrality, C_C – Closeness centraltity, C_C^- – Closeness centrality dle vstupních hran, C_D – Degree centrality, C_D^- – Degree centrality dle vstupních hran).

Tab. E.2: Uživatelé OSN seřazeni dle dosaženého místa pro jednotlivé metriky CM – 5% aproximace orientované sítě.

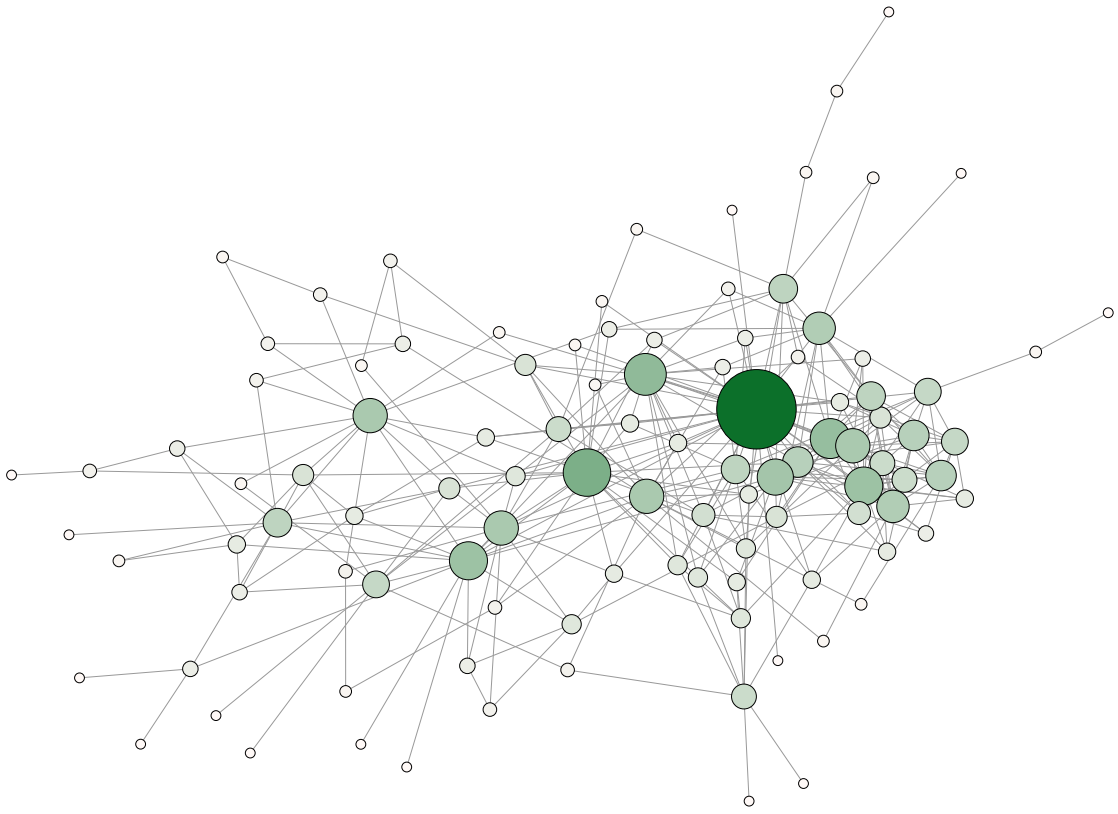
#	C_B	C_C^-	C_D^-
1	honzek098	r.hasek	kolousek2@email.cz
2	r.hasek	vivi.elien.186	avoldies
3	melly.19	honzek098	splichalpeta
4	sochurek.j	melly.19	vivi.elien.186
5	nezkrotny.dablicek.lucie.xd	x.x.radushka.princess.x.x	beruska.no1
6	durdil.d	kolousek2@email.cz	denny.hebdova
7	flamez10	lusie44	karollecka
8	vivi.elien.186	karollecka	italfllexo
9	kolousek2@email.cz	durdil.d	honzek098
10	martinoof	rompears	jitka.pabyskova
11	vvladimirvanek	lucixxcka-lucik	lucixxcka-lucik
12	avoldies	nezkrotny.dablicek.lucie.xd	domi-stara
13	ashraf.shafeek	flamez10	myska122
14	zlatovlaska111	xxxluciiinkaxxx	r.hasek
15	rompears	nikollka22	zpewanda
16	lukas.klement@post.cz	bacio333	durdil.d
17	01pitrisek	adrika75	marcela-marsi
18	martin.karatista	asuss	jahodka444
19	zlatahela50@email.cz	jahodka444	petra.dark
20	italfllexo	elvislu	rompears

Tab. E.3: Uživatelé OSN seřazeni dle dosaženého místa pro jednotlivé metriky CM – 20% aproximace orientované sítě.

#	C_B	C_C^-	C_D^-
1	honzek098	r.hasek	kolousek2@email.cz
2	r.hasek	vivi.elien.186	avoldies
3	melly.19	rompears	splichalpeta
4	durdil.d	melly.19	vivi.elien.186
5	sochurek.j	durdil.d	beruska.no1
6	flamez10	x.x.radushka.princess.x.x	denny.hebdova
7	nezkrotny.dablicek.lucie.xd	honzek098	karollecka
8	martinoof	lucixxcka-lucik	italfllexo
9	vivi.elien.186	kolousek2@email.cz	honzek098
10	kolousek2@email.cz	karollecka	jitka.pabyskova
11	vvladimirvanek	lusie44	lucixxcka-lucik
12	rompears	flamez10	domi-stara
13	ashraf.shafeek ashraf.shafeek	lucinnkaa16	myska122
14	avoldies	nezkrotny.dablicek.lucie.xd	r.hasek
15	martin.karatista	nikollka22	zpewanda
16	lukas.klement@post.cz	bacio333	durdil.d
17	01pitrisek	xxxluciiinkaxxx	marcela-marsi
18	zsidlo	jjjjoenick	jahodka444
19	asuss	asuss	petra.dark
20	italfllexo	horalka5	rompears

F Jádro sítě

V této příloze naleznete vizualizaci a několik informací o jádru sítě. To tvoří uživatelé, kteří mají více než 500 přátel nebo je za přátele považuje více než 500 uživatelů. Příloha se váže ke kapitole 6. Přes toto jádro je v síti přenášeno nejvíce informací. Je zajímavé, že téměř všechny vrcholy tvoří jednu komponentu (105 ze 123, tj. 85,4%). To ukazuje na velmi dobré propojení mezi centry. ASPL v této podsíti (pro neorientovaný model) je 3 a poloměr sítě 8. Pro orientovanou síť je pak ASPL 4 a poloměr sítě 10. Průměrný stupeň vrcholu je 5,2. Hustota jádra je 4,3E-2 pro síť neorientovanou a 2,1E-2 pro síť orientovanou. Je tedy výrazně hustší než síť tvořena všemi uživateli. Nejcentrálnějším uživatelem v jádře je dle všech zkoumaných metrik CM (C_D, C_C a C_B) uživatelka *nezkrotny.dablicek.lucie.xd*.



Obr. F.1: Největší komponenta jádra sítě – vrcholy jsou obarveny a zvětšeny dle metriky C_D (vytvořeno pomocí programu Gephi).